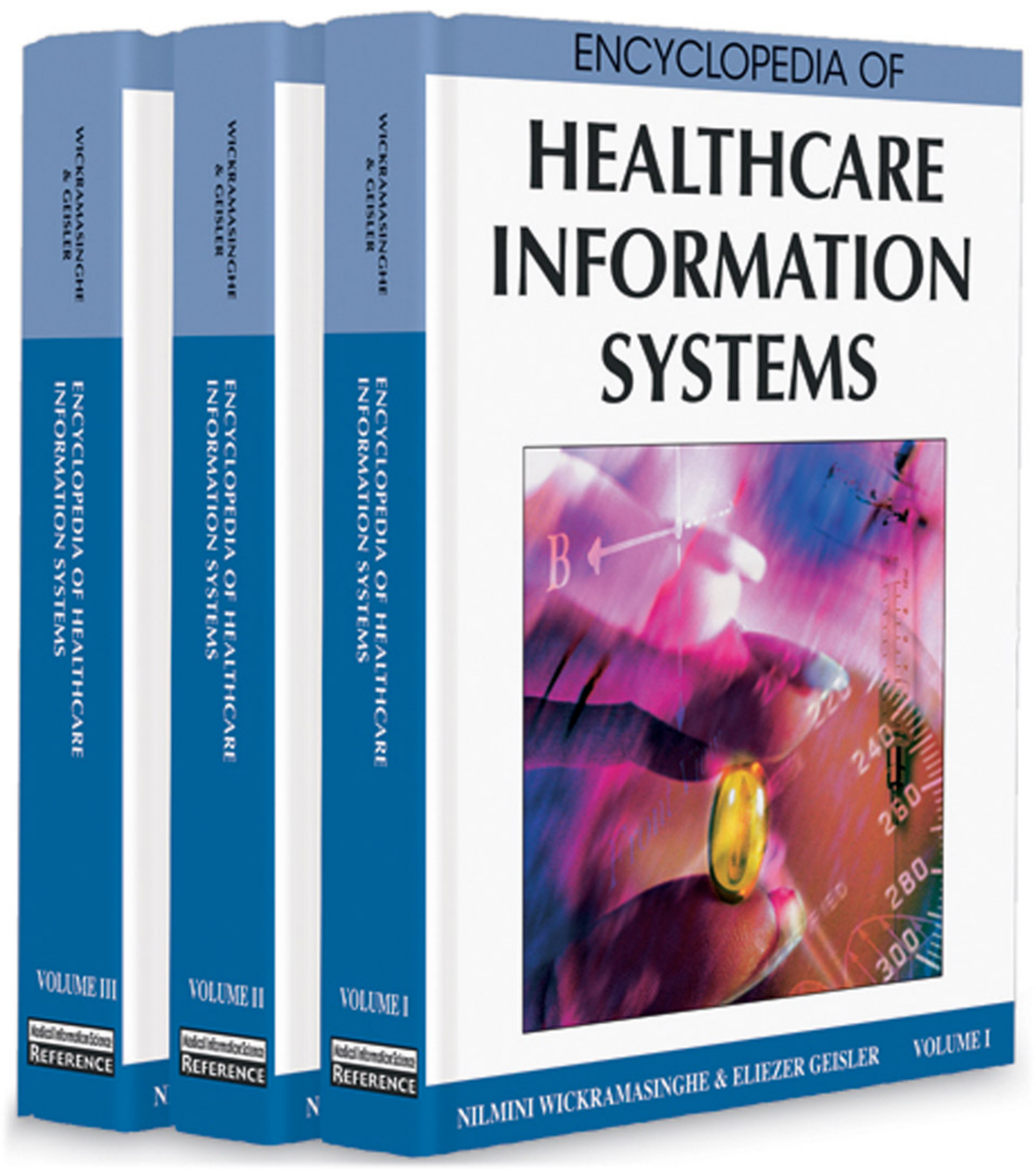

ENCYCLOPEDIA OF
HEALTHCARE
INFORMATION
SYSTEMS
NILMINI WICKRAMASINGHE & ELIEZER GEISLER
VOLUME I
WICKRAMASINGHE & GEISLER
ENCYCLOPEDIA OF HEALTHCARE INFORMATION SYSTEMS
VOLUME I
VOLUME II
VOLUME III
REFERENCE

Encyclopedia of Healthcare Information Systems

Nilmini Wickramasinghe
Illinois Institute of Technology, USA

Eliezer Geisler
Illinois Institute of Technology, USA

Medical Information Science Reference
MEDICAL INFORMATION SCIENCE REFERENCE
Hershey • New York

Acquisitions Editor:	Kristin Klinger
Development Editor:	Kristin Roth
Senior Managing Editor:	Jennifer Neidig
Managing Editor:	Jamie Snavely
Assistant Managing Editor:	Carole Coulson
Copy Editor:	Laura Kochanowski, Jeannie Porter, and Sue Vander Hook
Typesetter:	Jeff Ash and Sean Woznicki
Cover Design:	Lisa Tosheff
Printed at:	Yurchak Printing Inc.

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue, Suite 200
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com/reference

and in the United Kingdom by
Information Science Reference (an imprint of IGI Global)
3 Henrietta Street
Covent Garden
London WC2E 8LU
Tel: 44 20 7240 0856
Fax: 44 20 7379 0609
Web site: http://www.eurospanbookstore.com

Library of Congress Cataloging-in-Publication Data

Encyclopedia of healthcare information systems / Nilmini Wickramasinghe and Eliezer Geisler, editors.
p. ; cm.
Includes bibliographical references.
Summary: "This book provides an extensive and rich compilation of international research, discussing the use, adoption, design, and diffusion of information communication technologies (ICTs) in healthcare, including the role of ICTs in the future of healthcare delivery; access, quality, and value of healthcare; nature and evaluation of medical technologies; ethics and social implications; and medical information management"--Provided by publisher.
ISBN 978-1-59904-889-5 (h/c)
1. Medical informatics--Encyclopedias. I. Wickramasinghe, Nilmini. II. Geisler, Eliezer, 1942-
[DNLM: 1. Information Systems--Encyclopedias--English. W 13 E5523 2008]
R858.E52 2008
651.5'04261--dc22
2007047456

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

All work contributed to this encyclopedia set is original material. The views expressed in this encyclopedia set are those of the authors, but not necessarily of the publisher.

If a library purchased a print copy of this publication, please go to http://www.igi-global.com/agreement for information on activating the library's complimentary electronic access to this publication.

postera crescam laude

For my family, naturally – NW

To my grandchildren: Tyler, Max, Jack and Grace – Elie Geisler

Editorial Advisory Board

List of Contributors

Contents
by Volume

VOLUME I

VOLUME II

VOLUME III

Foreword

From a practitioner's point of view, this encyclopedia offers a wealth of academic rigor in regard to the use of information technology in both clinical and health care administrative settings. Many people working in the field find they encounter challenges when using technology with the goal to improve patient outcomes. One key challenge is finding ways to financially support research, develop, implement, and sustain new health care delivery processes. There has been some success achieved in small local settings and in countries where there is an inherent culture for collaboration. But it seems that large macro national and international achievements are still very allusive. Unlike any other market sector, health care is one of the most universal. This is evident with its demanding and globally accepted academic standards. This leads one to believe that health care is a perfect candidate to continuously deliver world wide successes. So why is it such a challenge to realize large-scale improvements in health care using technology?

When trying to implement new health care technology initiatives, the financial challenge may be a symptom of a much deeper problem. Upon further analysis, one possibility is there is a gap between practitioners and researchers. Both these players have a vital role in transitioning an idea to realization in the use of technology in health care. A researcher provides the academic rigor and peer review to achieve consensus, whereas a practitioner uses experience to implement these ideas in the field and supply quality data to support enhanced patient outcomes. Through their joint efforts, a unique core competency can be formalized and designed for the ever growing knowledge-based economy, with the hope that these endeavors will produce effective and efficient enhancements in the health system; a system that supports physicians, nurses, and other health care givers in patient care without access or cost barriers. Removing any gaps between research and its practice may be the catalyst necessary to produce results that achieve international acceptance and justify funding for worldwide execution.

Building new partnerships between researchers and practitioners may be the beginning of a new era in applying technology in health care. New linkages can be formed where both groups are working together on the same project. The researcher views it as a data collection project to support their studies and the practitioner views the project to support their technology implementations. Success is measured on how quickly the partnership can achieve a positive and global impact in health care.

In addition to offering a wealth of knowledge on health care information systems for researchers to share ideas, this encyclopedia is an excellent resource for practitioners potentially seeking new partnerships. Using this publication, practitioners have a wonderful opportunity to read, discover, and contact researchers that can help accelerate global achievements in applying health care information systems.

In summary, the encyclopedia can be viewed as a starting point to help form new information technology partnerships that support the growth of the knowledge economy, remove barriers from care giving, and enhance health care delivery for the most important person: the patient.

Steven Goldberg
President
INET International Inc.
Canada

Preface

Health care is an important industry that touches most, if not all of us at some time in our lives. Health care is noted for using leading edge technologies and embracing new scientific discoveries to enable better cures for diseases and better means to enable early detection of most life threatening diseases. However, the health care industry globally, and in the U.S. specifically, has been extremely slow to adopt technologies that focus on better practice management and administrative needs. Presently, health care is grappling with many challenges both nationally and globally, including escalating costs, a move to a preventative care environment, and a technology savvy patient with high expectations. The underlying goal for health care is to provide cost effective quality treatment, that is, realize its value proposition in this challenging environment. In order to do this, health care needs to maximize its information management and utilize information communication technologies (ICTs) prudently as has been noted by leaders of U.S. and the EU as well as leading bodies that focus on global health care issues and policy. During his State of the Union Address in January 2004, President George Bush affirmed the intention of the government to emphasize the role of technology in administration and delivery of health care in the United States (Bush, 2004) and allocated over $200 million for research and development of information technologies in health care. Similar sentiments have been voiced by the European leaders (Global Medical Forum Foundation, 2005; The Oslo Declaration on Health, Dignity and Human Rights, 2003) and the World Health Organization (E-Health in the Eastern Mediterranean, 1997; World Health Organization, 1998). Both European and U.S. authorities define their initiatives primarily in terms of medical information technology centering on computerized patient record (CPR) or, in more acceptable parlance, the HER electronic health record (Brailer & Terasawa, 2003). WHO's platform statement (World Health Organization, 1998) speaks of "health telematics policy," an all inclusive term that incorporates not only HER but essentially all health care services provided at a distance and based on the use of information communications technologies (ICT).

It is useful to think of the major challenges facing today's health care organizations in terms of the categories of demographics, technology, and finance. Demographic challenges are reflected by longer life expectancy and an aging population; technology challenges include incorporating advances that keep people younger and healthier; and finance challenges are exacerbated by the escalating costs of treating everyone with the latest technologies. Health care organizations can respond to these challenges by focusing on three key solution strategies; namely (1) access—caring for anyone, anytime, anywhere; (2) quality—offering world class care and establishing integrated information repositories; and (3) value—providing effective and efficient health care delivery. These three components are interconnected such that they continually impact on the other and all are necessary to meet the key challenges facing health care organizations today. Moreover, it is at the confluence of these key solution strategies that the role of ICTs in health care becomes of particular significance.

Access to health care for anyone, anywhere, and at any time is a goal that is very far from being achieved. ICT has a crucial role to play in improving access to health care. Although ICT cannot by itself perform such actions as diagnosing a medical condition, it can provide low-cost communication between physician and patient that makes a diagnosis possible and innovations could dramatically improve the effectiveness of such remote access as well as address minimal invasive surgery. Other aspects of access are less obvious, but are critical to controlling quality and costs of health care. These include timely access to medical records and universal accessibility to data among disparate information systems and medical devices (AHRQ, 2003; PITAC, 2001). In each case, improvements in ICT design and standards can dramatically increase access.

Quality in health care has six key goals (Wickramasinghe, Schaffer, & Geisler, in press):

1. **Safety:** Avoiding injuries to patients from the care that is intended to help them
2. **Effectiveness:** Providing services based on scientific knowledge to all who could benefit and refraining from providing services to those who will not benefit(i.e., avoiding under-use and overuse)

3. **Patient-centered:** Providing care that is respectful of and responsive to individual patient preferences, needs, and values and ensuring that patient values guide all clinical decisions
4. **Timeliness:** Reducing waiting and sometimes harmful delays for both those who receive care and those who give care
5. **Efficiency:** Avoiding waste
6. **Equitability:** Providing care that does not vary in quality based on personal characteristics.

These quality aims will be negatively impacted by poor information quality, flow, and integrity. Conversely, higher quality, flow, and integrity of information can help to reduce the large number of medical errors that currently permeate the health care system (Geisler, Lewis, Nayar, & Prabhaker, 2003; Moore & Wesson, 2002).

Value of health care incorporates the overarching goal of increased productivity, but this goal has distinct aspects depending upon one's perspective, and only some of these aspects are primarily monetary. To patients, the value proposition may be enhanced by reducing time spent in obtaining health care, lowering stress and worry, and increasing satisfaction with the experience. To physicians and clinical support personnel, value may be enhanced by the ability to remotely access medical records and monitor prescribed regimens and patient symptoms. To managers and investors, value may be enhanced by lowering costs of delivering health care in ways that do not compromise access or quality.

Because of the rapidly-decreasing cost structure of hardware, IT is now able to provide many functions that were not previously available, such as connectivity through mobile devices. Thus, IT innovations hold great potential for enhancing the value of health care to patients and stakeholders. Yet, it is not simply the introduction of IT per se that provides a good solution, but often it is necessary to combine IT with new management techniques such as total quality, knowledge management, and business process redesign to achieve success (Sharma, Wickramasinghe, & Gupta, 2004). When we look at health care, we can see that in order to enhance value, we must not simply focus on clinical care but also be cognizant of education, research, and administrative needs of this industry.

This encyclopedia provides an extensive and rich compilation of various ICT initiatives and the role that ICT plays and will play in the future of health care delivery be it in terms of enhancing clinical delivery or enabling administrative activities. International experts address one or more of the areas of access quality and value and thereby represent ways in which health care delivery can be made superior and the health care industry can begin to address it's the major challenges it faces in the 21st Century so that ultimately the most important person in the web of health care players, the patient, can be confident of receiving high quality cost effective health care anytime anywhere.

Presently, to the best of our knowledge, no such comprehensive encyclopedia exists that focuses on ICTs in health care. Hence, this encyclopedia is truly unique and serves to fill an important void in the existing literature. Moreover, given the sentiments of both EU and U.S. leadership regarding investing in ICTs to support superior health care delivery, the timing of this encyclopedia is most opportune.

The purpose of our encyclopedia is to bring together articles by international experts pertaining to critical concepts about the use, adoption, design, and diffusion of ICTs in health care. Ultimately the role of ICTs in health care must address the challenges faced by today's health care environment. We have identified that any truly beneficial solution must focus on access, quality, and value, and thus the management of medical technologies must relate to addressing and facilitating one or more of these solution strategies. We believe that in order to do this in a meaningful and systematic fashion it is essential to focus on five important themes, namely:

a. Generation, adoption, and utilization of medical technologies. This thrust includes research on the organization, financial and managerial aspects of the process of need identification for medical technologies by health care organizations, and the acquisition, diffusion, utilization, updating, replacement, and resources allocation for such technologies.
b. The nature of medical technologies and innovation and the role it plays in health care delivery. This thrust focuses on studies of the generation of medical technologies in industry, universities and the government sectors, and the processes by which such technologies are marketed to the health care delivery sector and their consequent impact on resolving the challenges faced by health care globally.
c. Evaluation of medical technologies. Included in this thrust are evaluation, assessment, monitoring, and auditing of costs and benefits from medical technologies, in health care delivery organizations, and in the supporting industries, such as insurance, regulatory agencies, manufacturing of medical technology, and pharmaceuticals.
d. Ethics, social implications, and patient value. This thrust focuses on ethical considerations and the role of patients in the health delivery sector. The center conducts research on accessibility, availability, and value derived from health care delivery, and the role that medical technologies play in this regard for patients in general, the underserved and uninsured, and the emerging empowerment of patients.

e. Management of medical information and emerging technologies. This thrust focuses on studies of emerging technologies such as telemedicine, telehealth, computerized medical records, e-health, knowledge and knowledge management, knowledge in health care, and the future of medical informatics. Topics include diffusion, evaluation, economics, and applications of these technologies to the health care sector.

These themes also form the thrust for our own on going research endeavours at IIT's Centre for the Management of Medical Technology (CMMT; http://www.stuart.iit.edu/cmmt/). Contributing authors have focused their article(s) on one or more of the above five general themes or thrusts which taken together are naturally not exhaustive but do serve to underscore the key areas within health care delivery.

The significance of ICTs in health care cannot be understated. A recent survey by the Cleveland Clinic noted that health care is the number one or two key issue for U.S. citizens, ahead of employment, war, and the economy. Health care is also the number two topic searched by users of the Internet worldwide and there are over 100,000 commercial Web sites dedicated solely to health care. In both the U.S. and globally, health care cannot survive without the application and adoption of a whole spectrum of technologies and since health care impacts all of us, it becomes a priority to all readers, both health care professionals and general citizens, to develop an understanding and appreciation of the various possibilities for ICT use within the health care domain. We are confident that this encyclopedia will be used by all leading health care organizations throughout the U.S. and globally as well as be a vital part of any universities reference collection in the area of ICT use in health care. We encourage all individuals wanting to understand the role of technology in health care delivery and more especially the nuances pertaining to the management of medical technology to also use this as a useful reference. We trust that all readers will enjoy this compilation and become more enlightened on the subject.

The Editors

Nilmini Wickramasinghe and Elie Geisler, 2007

REFERENCES

AHRQ. (2003). National Health care Quality Report, Agency for Health care Research and Quality, US Department of Health and Human Services. Retrieved February 15, 2008, from http://www.qualitytools.ahrq.gov/qualityreport/download_report.aspx

Brailer, D.J., & Terasawa, A.B. (2003). *Use and adoption of computer-based patient records.* California Health care Foundation (pp. 1-42).

Bush, G.W. (2004). State of the Union Address, The White House. Retrieved February 15, 2008, from http://www.whitehouse.gov/news/release/2004/01/20040120-7.html

E-Health in the Eastern Mediterranean (1997). Retrieved February 15, 2008, from http://www.emro.who.int/his/ehealth/ehealthPlan.htm

Geisler, E., Lewis, D., Nayar, M., & Prabhaker, P. (2003). Information integrity and organizational performance: A model and research directions. In *Proceedings of the Third Hospital of the Future Conference*, Warwick, UK.

Global Medical Forum Foundation. (2005, April 4-6). GMF IV: New frontiers in health care. Retrieved February 15, 2008, from http://www.globalmedicalforum.org/gmf_iv/session3.asp

Moore, T., & Wesson, R. (2002). Issues regarding wireless patient monitoring within and outside the hospital. In *Proceedings of the Second Hospital of the Future Conference*, Chicago, USA.

The Oslo Declaration on Health, Dignity and Human Rights (2003, June 12-13). In *7th Conference of European Health Ministers,* Oslo, Norway. Retrieved February 15, 2008, from http://odin.dep.no/hod/norsk/aktuelt/taler/monister/042071-990209/dpk-bn.html

PITAC. (2001). *Transforming healthcare through information technology.* President's Information Technology Advisory Committee Panel on Transforming Health Care Report. Retrieved February 15, 2008, from http://www.hpcc.gov/pubs/pitac/pitac-hc-9feb01.pdf

Sharma, S., Wickramasinghe, N., & Gupta, J. (2004). Knowledge management in health care. In Wickramasinghe et al. (Eds.), *Creating knowledge-based healthcare organisation*. Hershey, PA: Idea Group Publishing.

Wickramasinghe, N., Schaffer, J., & Geisler, E. (in press). Realizing the value proposition for healthcare by incorporating KM strategies and data mining techniques with the use of information communication technologies. *Int. J. Health care Technology and Management.*

World Health Organization. (1998). A Health Telematics Policy in Support of WHO's Health-for-All Strategy for Global Health Development: Report of the WHO Group Consultation on Health Telematics.

Acknowledgment

It is not possible to successfully complete a project of this magnitude without the cooperation and contributions of many people. We are most grateful for all the efforts, input and support given to us by the contributors, reviewers, our advisory board and the staff at IGI Global. We would especially like to thank Mehdi Khosrow-Pour for inviting us to produce this magnum opus, Jan Travers for managing this project, and Kristin Roth for sustaining us through the process and keeping us to schedule. We also acknowledge Stuart School of Business, Illinois Institute of Technology for affording us the time to work on this project and are indebted to our colleagues and students for many stimulating discussions and assistance throughout the process. Finally, a special thanks to our families.

Nilmini Wickramasinghe and Elie Geisler

About the Editors

Nilmini Wickramasinghe (PhD; MBA; GradDipMgtSt; BSc Amus., piano; Amus. A, violin) researches and teaches in several areas within information systems, including knowledge management, e-commerce and m-commerce, and organizational impacts of technology with particular focus on the applications of these areas to health care and thereby effecting superior health care delivery. She is well published in all these areas with more than 60 referred scholarly articles and several books. In addition, she regularly presents her work throughout North America, as well as in Europe and Australasia. Dr. Wickramasinghe is the U.S. representative of the Health Care Technology Management Association (HCTM), an international network of scholars that focuses on critical health care issues and the role of technology within the domain of health care. She is the associate director of the Center for the Management of Medical Technologies (CMMT) and holds an associate professor position at the Stuart School of Business, Illinois Institute of Technology. In addition, Dr. Wickramasinghe is the editor-in-chief of two scholarly journals: *International Journal of Networking and Virtual Organisations* (IJNVO – www.inderscience.com/ijnvo) and *International Journal of Biomedical Engineering and Technology* (IJBET- www.inderscience.com/ijbet), where she was also the journal's founder. Both journals are published by InderScience.

Elie Geisler is a distinguished professor at the Stuart School of Business, Illinois Institute of Technology and director of the IIT Center for the Management of Medical Technology. He holds a doctorate from the Kellogg School at Northwestern University. Dr. Geisler is the author of nearly 100 papers in the areas of technology and innovation management, the evaluation of R&D, science and technology, and the management of health care and medical technology. He is the author of several books, including: *Managing the Aftermath of Radical Corporate Change* (1997), *The Metrics of Science and Technology* (2000), also translated into Chinese, *Creating Value with Science and Technology* (2001) and *Installing and Managing Workable Knowledge Management Systems* (with Rubenstein, 2003). His forthcoming books are: *Knowledge and Knowledge Systems: Learning from the Wonders of the Mind*, and a textbook: *Knowledge Management Systems: Theory and Cases* (with Wickramasinghe). He consulted for major corporations and for many U.S. federal departments, such as Defense, Agriculture, Commerce, EPA, Energy, and NASA. Dr. Geisler is the cofounder of the annual conferences on the Hospital of the Future, and the Health Care Technology and Management Association, a joint venture of several universities in 10 countries. He serves on various editorial boards of major journals. His current research interests include the nature of knowledge management in general and in complex systems, such as health care organizations.

3D and 4D Medical Image Registration Combined with Image Segmentation and Visualization

Guang Li
National Cancer Institute, USA

Deborah Citrin
National Cancer Institute, USA

Robert W. Miller
National Cancer Institute, USA

Kevin Camphausen
National Cancer Institute, USA

Boris Mueller
Memorial Sloan-Kettering Cancer Center, USA

Borys Mychalczak
Memorial Sloan-Kettering Cancer Center, USA

Yulin Song
Memorial Sloan-Kettering Cancer Center, USA

INTRODUCTION

Image registration, segmentation, and visualization are three major components of medical image processing. Three-dimensional (3D) digital medical images are three dimensionally reconstructed, often with minor artifacts, and with limited spatial resolution and gray scale, unlike common digital pictures. Because of these limitations, image filtering is often performed before the images are viewed and further processed (Behrenbruch, Petroudi, Bond, et al., 2004). Different 3D imaging modalities usually provide complementary medical information about patient anatomy or physiology. Four-dimensional (4D) medical imaging is an emerging technology that aims to represent patient motions over time. Image registration has become increasingly important in combining these 3D/4D images and providing comprehensive patient information for radiological diagnosis and treatment.

3D images have been utilized clinically since computed tomography (CT) was invented (Hounsfield, 1973). Later on, magnetic resonance imaging (MRI), positron emission tomography (PET), and single photon emission computed tomography (SPECT) have been developed, providing 3D imaging modalities that complement CT. Among the most recent advances in clinical imaging, helical multislice CT provides improved image resolution and capacity of 4D imaging (Pan, Lee, Rietzel, & Chen, 2004; Ueda, Mori, Minami et al., 2006). Other advances include mega-voltage CT (MVCT), cone-beam CT (CBCT), functional MRI, open field MRI, time-of-flight PET, motion-corrected PET, various angiography, and combined modality imaging, such as PET/CT (Beyer, Townsend, Brun et al., 2000), and SPECT/CT (O'Connor & Kemp, 2006). Some preclinical imaging techniques have also been developed, including parallel multichannel MRI (Bohurka, 2004), Overhauser enhanced MRI (Krishna, English, Yamada et al., 2002), and electron paramagnetic resonance imaging (EPRI) (Matsunoto, Subramanian, Devasahayam et al., 2006).

Postimaging analysis (image processing) is required in many clinical applications. Image processing includes image filtering, segmentation, registration, and visualization, which play a crucial role in medical diagnosis/treatment, especially in the presence of patient motion and/or physical changes. In this article, we will provide a state-of-the-art review on 3D/4D image registration, combined with image segmentation and visualization, and its role in image-guided radiotherapy (Xing, Thorndyke, Schreibmann et al., 2006).

BACKGROUND

3D/4D Medical Imaging

A 3D medical image contains a sequence of parallel two-dimensional (2D) images representing anatomic or physiologic information in 3D space. The smallest element of a 3D image is a cubic volume called voxel. A 4D medical image contains a temporal series of 3D images. With a subsecond time resolution, it can be used for monitoring respiratory/cardiac motion (Keall, Mageras, Malter et al., 2006).

Patient motion is always expected: faster motion relative to imaging speed causes a blurring artifact; whereas slower motion may not affect image quality. A multislice CT scanner provides improved spatial and temporal resolution (Ueda et al., 2006), which can be employed for 4D imaging (Pan et al., 2004). Progresses in MRI imaging have also been reported, including parallel multichannel MRI (Bodurka, Ledden, van Gelderen et al., 2004).

Because PET resolution and speed are limited by the physics and biology behind the imaging technique, some motion suppression techniques have been developed clinically, including patient immobilization (Beyer, Tellmann, Nickel, & Pietrzyk, 2005), respiratory gating (Hehmeh, Erdi, Pan et al., 2004), and motion tracking (Montgomery, Thielemans, Mehta et al., 2006). Motion tracking data can be used to filter the imaging signals prior to PET image reconstruction for reliable motion correction. Motion blurring, if uncorrected, can reduce registration accuracy. Visual-based volumetric registration technique provides blurring correction (filtering) before registration, by defining the PET volume with reference to the CT volume, causing blurred PET surface voxels to be rendered invisible (Li, Xie, Ning et al., 2007).

Image Segmentation and Visualization

Medical image segmentation defines regions of interest used to adapt image changes, study image deformation, and assist image registration. Many methods for segmentation have been developed including thresholding, region growing, clustering, as well as atlas-guided and level sets (Pham, Xu, & Prince, 2000; Suri, Liu, & Singh et al., 2002). Atlas-guided methods are based on a standard anatomical atlas, which serves as an initial point for adapting to any specific image. Level sets, also called active contours, are geometrically deformable models, used for fast shape recovery. Atlas-based level sets have been applied clinically for treatment planning (Lu, Olivera, Chen et al., 2006a; Ragan, Starkschall, McNutt et al., 2005;) and are closely related to image registration (Vemuri, Ye, & Chen, et al., 2003). Figure 1 shows automatic contours. Depending on how the 3D image is segmented, it can be either 2D-based or 3D-based (Suri, Liu, Reden, & Laxminarayan, 2002).

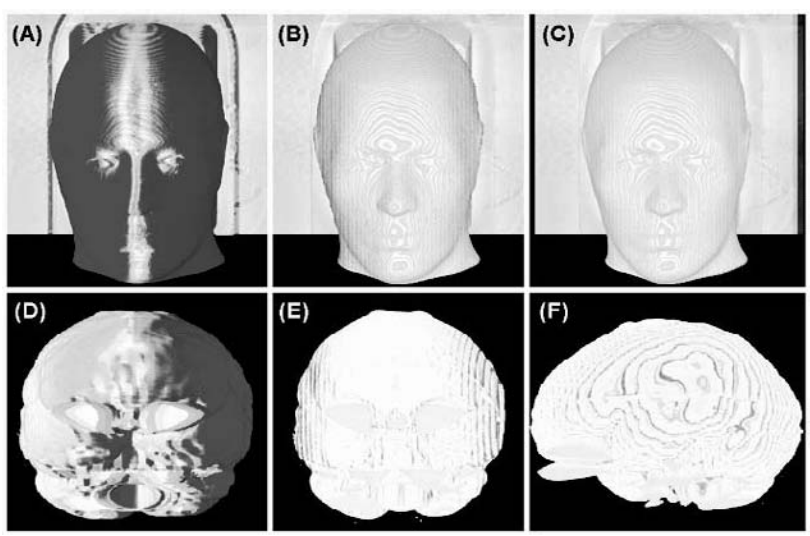
3D medical image visualization has been increasingly applied in diagnosis and treatment (Salgado, Mulkens, Bellinck, & Termote, 2003), whereas 2D-based visualization is predominantly applied clinically. Because of the demand on computing power, real-time 3D image visualization is supported by specialized graphics hardware (Terarecon, Inc.) (Xie, Li, Ning et al., 2004) or high-end consumer graphics processors (Levin, Aladl, Germanos, & Slomak, 2005). 3D image visualization has been applied to registration of four imaging modalities with improved spatial accuracy (Li, Xie, Ning et al., 2005; Li et al., 2007). Figure 2 shows 3D volumetric image registration using external and internal anatomical landmarks.

Rigid Image Registration

Rigid image registration assumes a motionless patient such that the underlying anatomy is identical in different imaging modalities for alignment. Three approaches to rigid registration are: coordinate-based, extrinsic-based, and intrinsic-based (Maintz & Viergever, 1998). Coordinate-based registration is performed by calibrating the coordinate system to produce “co-registered” images. Multimodality scanners, such as PET/CT and SPECT/CT, are typical examples.

Figure 1. Orthogonal 2D-views of CT images and comparison of automatic recontours (solid-lines) and manual contours (dash-lines) in different phases (A&B) of a radiotherapeutic treatment (courtesy of Dr. Weiguo Lu)

Figure 2. 3D-views of CT (A, B, and C: head) and MR (D, E, and F: segmented brain) phantom images. The homogeneity of color distributed on an anatomic landmark is used as the registration criterion. From (A) to (C), three images (red, green, and blue) are approaching registration with shifts from 5.0mm to 0.5mm and 0.0mm, respectively. From (D) to (E and F), four images (grey, blue, green, and red) are 5.0mm apart from each other (D) and registered in front (E) and side (F) views.



Extrinsic-based image registration relies on the alignment of extrinsic objects placed in/on a patient invasively/noninvasively. Such objects can be fiducials or frames that are visible in all imaging modalities and serve as local coordinate markers (sets of points) for rigid registration. Examples are gold seeds for prostate localization in radiotherapy and head frame for stereotactic radiosurgery.

Intrinsic-based image registration uses a patient's anatomy (anatomic landmarks, segmented geometries, or intact voxels) as the registration reference. Alignment of visual landmarks or segmented geometries requires user interaction, so the registration is manual or semi-automatic. The statistical similarity of the intact voxels (grayscale) of two images, such as mutual information (Viola & Wells, 1995), has been widely used for fully automated registration (Pluim, Maintz, & Viergever, 2003).

Automatic image registration requires three key elements: a metric function, a transformation, and an optimization process. One common voxel-based image registration uses mutual information as the metric, a rigid or nonrigid transformation and a maximization algorithm. Recently, the homogeneity of color distributed on a volumetric landmark has been used as quantitative metric, assisted by the ray-casting algorithm in 3D visualization (Li et al., 2007). Figures 3 and 4 show clinical examples using the 3D visualization-based registration technique.

Deformable Image Registration

Deformable image registration contains a nonrigid transformation model that specifies the way to deform one image to match another. A rigid image registration is almost always performed to determine an initial position using rigid transformation with six variables (3 translations and 3 rotations). For nonrigid transformation, the number of variables will increase dramatically, up to three times the number of voxels. Common deformable transformations are spline-based with control points, the elastic model driven by image similarity, the viscous fluid model with region growth, the finite element model using rigidity classification, the optical flow with motion estimation, and free-form deformation (Chi, Liang, & Yan, 2006; Crum, Hartkens, & Hill, 2004; Lu, Olivera, Chen et al., 2006b).

The image similarity measures are ultimately the most important criteria for determining the quality of registration. They can be feature-based and voxel-based (Maintz & Viergever, 1998). The former is usually segmentation/classification based, adapting changes in shape of anatomical landmarks, while the latter is based on statistical criteria for intensity pattern match-

Figure 3. 3D-views of before (A, B, and C) and after (D, E, and F) 3D volumetric image registration, using homogeneity color distribution as registration criterion. Voluntary patient head movement is corrected in three MR images, T1 (green), FLAIR (red), and T2 (blue), which are acquired in the same scanner with time interval of 3 and 20 minutes.

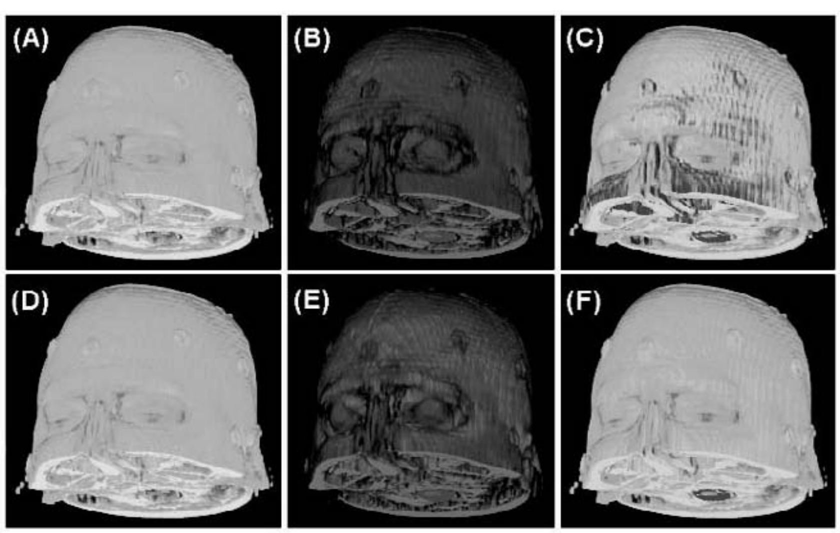


Figure 4. 3D-views of before (A, B, and C) and after (D, E, and F) rigid volumetric image registration of PET/CT images, correcting patient movement.

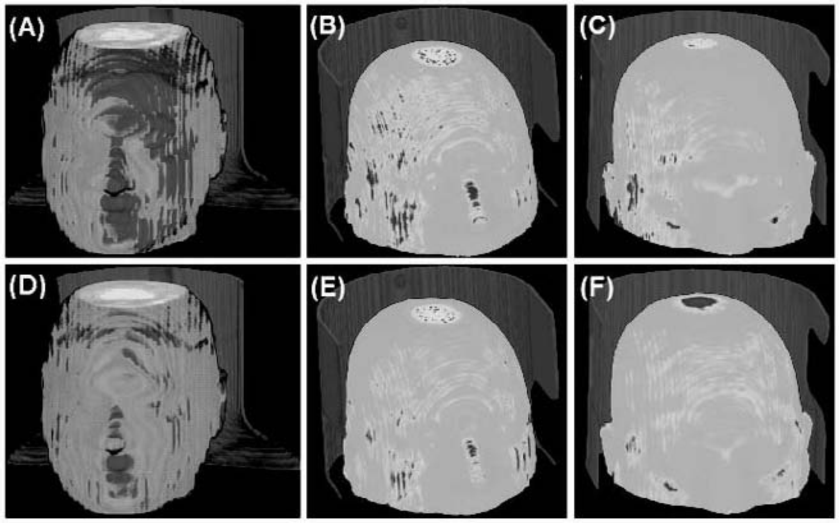
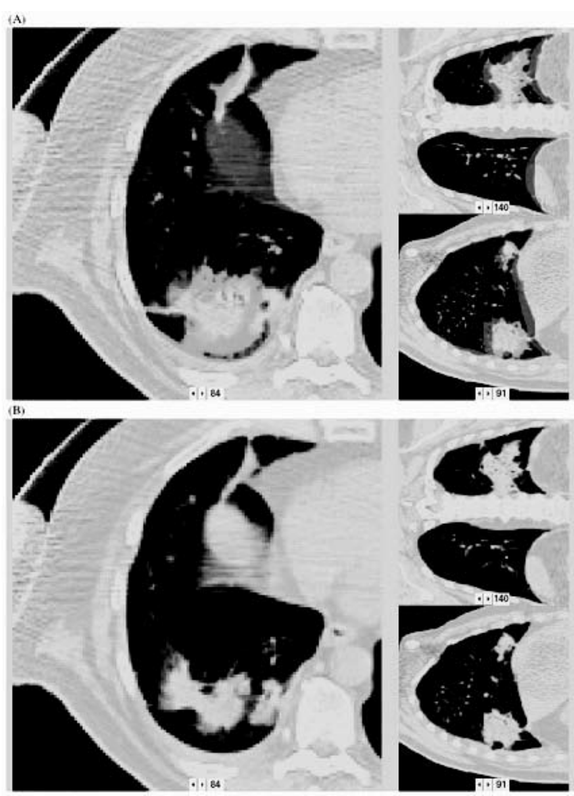


Figure 5. Orthogonal 2D-views of before (A) and after (B) deformable image registration of two 3D images (red and green) in a 4D CT image (courtesy of Dr. Weiguo Lu)



ing, including mutual information. Most deformable registrations are automated. Combining the two methods can improve registration accuracy, reliability, and/or performance (Hellier & Barillot, 2003; Liu, Shen, & Davatizikos, 2004; Wyatt & Noble, 2003). Figure 5 shows one example of deformable registration.

A Challenge from 3D/4D Conformal Radiotherapy: Deformable Image Registration

Broadened Concept of 4D Medical Imaging

The 4D imaging concept has been broadened to cover various time resolutions. The common 4D image has subsecond temporal resolution (Pan et al., 2004), while a series of 3D images, reflecting patient changes over a longer time span, should be also qualified as a 4D image with sufficient resolution to assess slower changes, including tumor growth/shrinkage and weight gain/loss during a course of treatment.

Registration of a 3D image to a 4D image involves a series of deformable registration. Because patient motion/change is inevitable, deformable image registration is the key to combining these images for clinical use. Clinically, MVCT images can be acquired daily and used for patient daily setup via rigid registration to the reference planning image, assuming minimal patient changes. Within a treatment, 4D CT imaging has shown dramatic anatomical changes during respiration (Keall et al., 2006). Image-guided frameless cranial and extra-cranial stereotactic radiosurgery has been performed clinically (Gibbs, 2006). Rigid image registration provides the best current solution to these clinical applications. Ultimately, deformable image registration will improve the registration accuracy substantially (Lu et al., 2006b), permitting highly conformal 3D/4D radiotherapy (Barbiere, Hanley, Song et al., 2007; Mackie, Kapatoes, Ruchala et al., 2003).

Challenges in Deformable Image Registration

For deformable image registration, the underlying anatomy changes, therefore voxel mapping among images is a challenge. First, the deformable transformation handles a large number of positioning variables that must be determined for every voxels within the anatomy. This can be abstracted as a multiple variable optimization problem in mathematics, limiting the performance of deformable image registration for many years (Crum, 2004). Second, the deformable registration is extremely difficult to be validated as there is lack of absolutes with respect to the location of corresponding voxels. Therefore, the accuracy and reliability of deformable registration should be evaluated on a case specific basis (Sanchez Castro, Pollo, Meuli et al., 2006; Wang, Dong, O'Daniel et al., 2005).

Regardless the limitations above, progress has been made by combining image registration and segmentation/classification to provide intrinsic simplification and cross verification. It remains a challenge, however, to develop a fully automated deformable registration algorithm because image segmentation often requires human interaction.

Deformable image registration is generally a "passive" mapping process. It does not anticipate how patient anatomy might deform. An example is whether superficial 3D contour information detected by a real-time infrared camera can be used to predict the motion of internal organs (Rietzel, Rosenthal, Gierga et al., 2004). Anatomically, the correlation between superficial and internal organ motion should exist, although as a complex relationship. Therefore, an anatomic model based image registration with motion estimation can provide an "active" mapping process, but is beyond the current scope of deformable image registration.

Gaps between Frontier Research and Clinical Practice

Despite of the advances in 3D and 4D imaging and image registration, 2D-based rigid registration techniques are predominantly used in the clinic; although automatic rigid registration methods exist in most commercial treatment planning software. Two reasons are primarily responsible for this disconnect: First, the user must visually verify the final registration using the 2D-based visualization tools available for image fusion in most commercial software. Second, most clinical images have some degree of pre-existing deformation so that automatic rigid registration can prove unreliable but manual methods allow the user to perform local organ registration. Some recent commercial software has recognized this problem and provides the option of selecting the region-of-interest to ease the deformation problem. This method, however, is only a partial solution to cope with image changes.

The gap between the clinical research and routine practice can be reduced by translational research and development. Recently, open-source medical image processing and visualization tool kits have become available for public use. Many recently published algorithms in medical image processing are implemented in a generic, object-oriented programming style, which permits reusability of such toolkits.

FUTURE TRENDS

3D rigid image registration will dominate clinical practice and will remain essential as more specialized complementary 3D imaging modalities become clinically relevant. Although the simplicity of automatic image registration is more attractive, manual image registration with 2D/3D visualization is irreplaceable because it permits incorporation of medical knowledge for verification and adjustment of the automatic registration results.

As awareness of the problems of the patient motion and anatomic changes increases, further research on 4D imaging and deformable registration will be stimulated to meet the clinical demands. Motion correction in the PET/CT and SPECT/CT will continue to improve the “coregistration” of these images. Interdisciplinary approaches are expected to offer further improvements for the difficult registration problem. With advances in hybrid registration algorithms and parallel computing, more progresses are expected, resulting in improved accuracy and performance.

CONCLUSION

Higher dimensional deformable image registration has become a focus of clinical research. The accuracy, reliability, and performance of 3D/4D image registration have been improved with assistance of image segmentation and visualization.

REFERENCES

Barbiere, J., Hanley, J., Song, Y., et al. (2007). Concise matrix analysis of point-based prostate targeting for intensity modulated radiation therapy. *Technol. Cancer Res. Treat., 6*(1), 1-10.

Behrenbruch, C.P., Petroudi, S., Bond, S., et al. (2004). Image filtering techniques for medical image post-processing: An overview. *Br. J. Radiol., 77*, S126-132.

Beyer, T., Tellmann, L., Nickel, I., & Pietrzyk, U. (2005). On the use of positioning aids to reduce misregistration in the head and neck in whole-body PET/CT studies. *J. Nucl. Med., 46*(4), 596-602.

Beyer, T., Townsend, D.W., Brun, T., et al. (2000). A combined PET/CT scanner for clinical oncology. *J. Nucl. Med., 41*(8), 1369-79.

Bodurka, J., Ledden, P.J., van Gelderen, P., et al. (2004). Scalable multichannel MRI data acquisition system. *Magn. Reson. Med., 51*, 165-171.

Chi, Y., Liang, J., & Yan, D. (2006). A material sensitivity study on the accuracy of deformable organ registration using linear biomechanical models. *Med. Phys., 33*(2), 421-433.

Crum, W.R., Hartkens, T., & Hill, D.L.G. (2004). Nonrigid image registration: Theory and practice. *Br. J. Radiol., 77*, S140-S153.

Gibbs, I.C. (2006). Frameless image-guided intracranial and extracranial radiosurgery using the Cyberknife robotic system. *Cancer Radiother, 10*(5), 283-287.

Hehmeh, S.A., Erdi, Y.E., Pan, T., et. al. (2004). Four-dimensional (4D) PET/CT imaging of the thorax. *Med. Phys., 31*(12), 3179-3186.

Hellier, P., & Barillot, C. (2003). Coupling dense and landmark-based approaches for nonrigid registration. *IEEE Trans. Med. Imaging., 22*, 217-227.

Hounsfield, G. (1973). Computerized transverse axial scanning (tomography): Part I. Description of system. *Br. J. Radiol., 46*, 1016.

Keall, P.J., Mageras, G.S., Malter, J.M., et al. (2006). The management of respiratory motion in radiation oncology report of AAPM Task Group 76. *Med. Phys., 33*(10), 3874-3900.

Krishna, M.C., English, S., Yamada, K., et al. (2002). Overhauser enhanced magnetic resonance imaging for tumor oximetry: Coregistration of tumor anatomy and tissue oxygen concentration. *Proc. Natl. Acad. Sci., 99*(4), 2216-2221.

Levin, D., Aladl, U., Germano, G., & Slomak, P. (2005). Techniques for efficient, real-time, 3D visualization of multi-modality cardiac data using consumer graphics hardware. *Compu. Med. Imaging Graph., 29,* 463-475.

Li, G., Xie, H., Ning, H., et. al. (2005). A novel 3D volumetric voxel registration technique for volume-view-guided image registration of multiple imaging modalities. *Int. J. Rad. Oncol. Biol. Phys., 63*(1), 261-273.

Li, G., Xie, H., Ning, H., et. al. (2007). Registering molecular imaging information into anatomic images with improved spatial accuracy. In *Proc. Int. Symp. Biomed. Imaging, IEEE Signal Processing Society* (pp.1144-1147).

Liu, T.M., Shen D.G., & Davatizikos, C. (2004). Deformable registration of cortical structures via hybrid volumetric and surface warping. *NeuroImage, 22*, 1790-1801.

Lu, W., Olivera, G.H., Chen, Q., et. al. (2006a). Automatic re-contouring in 4D radiotherapy. *Phys. Med. Biol., 51*, 1077-1099.

Lu, W., Olivera, G.H., Chen, Q., et. al. (2006b). Deformable registration of the planning image (kVCT) and the daily images (MVCT) for adaptive radiation therapy. *Phys. Med. Biol., 51,* 4357-4374.

Mackie, T.R., Kapatoes, J., Ruchala, K., et. al. (2003). Image guided for precise conformal radiotherapy. *Int. J. Rad. Oncol. Biol. Phys., 56*(1), 89-105.

Maintz, J.B.A., & Viergever, M.A. (1998). A survey of medical image registration. *Med. Image Anal., 2*(1), 1-36.

Matsunoto, K., Subramanian, S., Devasahayam, N., et al. (2006). Electron paramagnetic resonance imaging of tumor hypoxia: Enhanced spatial and temporal resolution for in vivo pO_2 determination. *Magn. Reson. Med., 55,* 1157-1163.

Montgomery, A.J., Thielemans, K., Mehta, M.A., et al. (2006). Correction of head movement on PET studies: Comparison of methods. *J. Nucl. Med., 47,* 1936-1944.

O'Connor, M.K., & Kemp, B.J. (2006). Single-photon emission computed tomography/computed tomography: Basic instrumentation and innovations. *Semin. Nucl. Med. 36*(4), 258-266.

Pan, T., Lee, T.-Y., Rietzel, E., & Chen, G.T. (2004). 4D-CT imaging of a volume influenced by respiratory motion on multi-slice CT. *Med. Phys., 31*(2), 333-340.

Pham, D.L., Xu, C., & Prince, J.L. (2000). Current methods in medical image segmentation. *Annu. Rev. Biomed. Eng., 2,* 315-237.

Pluim, J.P.W., Maintz, J.B.A., & Viergever, M.A. (2003). Mutual-information-based registration of medical images: A survey. *IEEE Trans. Med. Imaging, 22*(8), 986-1004.

Ragan, D., Starkschall, G., McNutt, T., et al. (2005). Semiautomated four-dimensional computed tomography segmentation using deformable models. *Med. Phys., 32*(7), 2254-2261.

Rietzel, E., Rosenthal, S.J., Gierga, D.P., et al. (2004). Moving targets: Detection and tracking of internal organ motion for treatment planning and patient set-up. *Radiother. Oncol., 73*(S2), S68-S72.

Salgado, R., Mulkens, T., Bellinck, P., & Termote, J.-L. (2003). Volume rendering in clinical practice: A pictorial review. *JBR-BTR, 86*(4), 215-220.

Sanchez Castro, F.J., Pollo, C., Meuli, R., et al. (2006). A cross validation study of deep brain stimulation targeting: From experts to atlas-based, segmentation-based and automatic registration algorithms. *IEEE Trans. Med. Imaging, 25*(11), 1440-1450.

Suri, J.S., Liu, K., Singh, S., et al. (2002). Shape recovery algorithms using level sets in 2-D/3-D medical imagery: A state-of-the-art review. *IEEE Trans. Inform. Tech. Biomed., 6*(1), 8-28.

Suri, J.S., Liu, K. Reden, L., & Laxminarayan, S. (2002). A review on MR vascular image processing: Skeleton versus nonskeleton approaches: Part II. *IEEE Trans. Inform. Tech. Biomed., 6*(4), 338-350.

Ueda, T., Mori, K., Minami, M., et al. (2006). Trends in oncological CT imaging: Clinical application of multidetector-row CT and 3D-CT imaging. *Int. J. Clin. Oncol., 11,* 268-277.

Vemuri, B.C., Ye, J., Chen, Y., & Leonard, C.M. (2003). Image registration via level-set motion: Applications to atlas-based segmentation. *Med. Image Anal., 7,* 1-20.

Viola, P. & Wells, III, W.M. (1995). Alignment by maximization of mutual information. In *Int. Conf. Computer Vision, IEEE Computer Society Press,* Los Alamitos, CA (pp. 16-23).

Wang, H., Dong, L., O'Daniel, J., et al. (2005). Validation of an accelerated demons algorithm for deformable image registration in radiation therapy. *Phys. Med. Biol., 50,* 2887-2905.

Wyatt, P.P., & Noble, J.A. (2003). MAP MRF joint segmentation and registration of medical images. *Med. Image Anal., 7,* 539-552.

Xie, H., Li, G., Ning, H., et al. (2004). 3D voxel fusion of multi-modality medical images in a clinical treatment planning system. In R. Long, S. Antani, D.J. Lee, B. Nutter, M. Zhang (Eds.), *Proc. Computer-Based Med. Sys., IEEE Computer Science Society* (pp. 48-53).

Xing, L, Thorndyke, B., Schreibmann, E., et al. (2006). Overview of image-guided radiation therapy. *Med. Dosim., 31*(2), 91-112.

KEY TERMS

3D Medical Imaging: A process of obtaining a 3D volumetric image composed of multiple 2D images, which are computer reconstructed using a mathematical "back-projection" operation to retrieve pixel data from projected image signals through a patient, detected via multichannel detector arrays around the patient.

4D Medical Imaging: A process of acquiring multiple 3D images over time prospectively or retrospectively, so that patient motions and changes can be monitored and studied.

Imaging Modality: A type of medical imaging technique that utilizes a certain physical mechanism to detect patient internal signals that reflect either anatomical structures or physiological events.

Image Processing: A computing technique in which various mathematical operations are applied to images for image enhancement, recognition, or interpretation, facilitating human efforts.

Image Registration: A process of transforming a set of patient images acquired at different times and/or with different modality into the same coordinate system, mapping corresponding voxels of these images in 3D space, based on the underlying anatomy or fiducial markers.

Image Segmentation: A process in which an image is partitioned into multiple regions (sets of pixels/voxels in 2D/3D) based on a given criterion. These regions are nonoverlapping, homogeneous with respect to some characteristics such as intensity or texture. If the boundary constraint of the region is removed, the process is defined as classification.

Image Visualization: A process of converting (rendering) image pixel/voxel into 2D/3D graphical representation. Most computers support 8-bit (256) grayscale display, sufficient to human vision that can only resolve 32-64 grayscale. A common 12/16-bit (4096/65536 grayscales) medical image can be selectively displayed based on grayscale classification. Window width (display range in grayscale) and linear level function (center of the window width) are frequently used in adjusting display content.

The ABC Approach and the Feminization of HIV/AIDS in the Sub-Saharan Africa

Lynette Kvasny
The Pennsylvania State University, USA

Jing Chong
The Pennsylvania State University, USA

INTRODUCTION

The growing prevalence of HIV/AIDS infections among women in African nations south of the Sahara is a complex and pressing public health concern. In this article, we examine how HIV/AIDS prevention campaigns construct women as the new face of HIV/AIDS in Sub-Saharan Africa. We do so by providing a feminist analysis of the US Government's Abstain, Be faithful, and correct and consistent use of condoms (ABC) health campaign. President Bush's Emergency Plan for AIDS Relief is the largest commitment ever made by a single nation towards an international health initiative—a five-year, $15 billion approach to combating HIV/AIDS. The centerpiece of the prevention component of this plan is the ABC approach (Office of the United States Global AIDS Coordinator, 2005). Abstinence, according to this theory, should take precedence for people who are not in a relationship. Those who are in a relationship should remain faithful to their partners. And if the first two strategies fail for any reason, condoms should be used to prevent the transmission of HIV. Global AIDS Coordinator Randall Tobias endorsed a provision in U.S. law requiring that at least one-third of all U.S. assistance to prevent HIV/AIDS globally be reserved for "abstinence-until-marriage" programs. In effect, this makes "abstinence-until-marriage" advocacy the single most important HIV/AIDS prevention intervention of the U.S. government.

In our feminist analysis, we view this HIV/AIDS prevention campaign from a cultural perspective (Cheng, 2005; Sontag, 1990; Treichler, 1999; Waldby, 1996), rather than structural determinants of risk such as public policy, globalization, industrialization, and the economy**.** For us, the feminization of this disease operates as an epidemic of signification, which takes the gendered, raced, and sexualized body of women as its subject. Far from the "gay white men's disease" of the 1980s, HIV/AIDS is infecting and affecting women more than ever before. As the epidemic enters its third decade, women now account for half of the 42 million people living worldwide with HIV/AIDS. Of the 3.1 million people who died of AIDS in 2002, 1.2 million were women. In Sub-Saharan Africa, the transmission rate for women has surpassed that of men with six HIV-positive women for every five HIV-positive men (UNAIDS/WHO, 2004).

Women, the dominant subjects of AIDS discourses, are placed at risk by common systems of oppression such as gender, race, class, and social and spatial location. Through the health campaigns which are disseminated and reproduced through television, radio, newspapers, and more recently the internet, women are uniquely constructed by privileged "experts" from the West as consumable subjects. In the case of women in Sub-Sahara Africa, we found that health campaigns which feminize AIDS are rooted in largely hegemonic cultural images which portray women as vulnerable subjects under siege. Through our analysis, we problematize the ABC health campaign and its appropriateness for women in Sub-Saharan Africa.

BACKGROUND

Women in Sub-Saharan Africa have become the new face of HIV/AIDS . While calling attention to women may help to end their silent suffering, if not done sensitively, it may unwittingly reproduce a discourse that depicts Africa in largely pessimistic terms. Media images of Black children with emaciated bodies, impoverished communities facing environmental and epidemic catastrophes, and bare-breasted women standing besides grass huts are imprinted on the collective consciousness of citizens in the West. The internet provides a global forum for disseminating "afropessimism" through a broad range of communication channels, including televised and printed media reports, news outlets, medical journals, Web sites, press releases, and policy documents, among others.

Through these pessimistic portrayals, African women have come to serve as a convenient trope for signifying the worst of the global HIV/AIDS pandemic, a fate to be both feared and avoided by other nations. The human suffering resulting from HIV/AIDS in Sub-Saharan Africa has become entrenched in the psyche of the general U.S. population. For instance, the Kaiser Family Foundation's *2004 Survey of Americans on HIV/AIDS* found that more than eight in 10 (83%) correctly acknowledged Africa as the part of the world that has been hardest hit with HIV/AIDS. However, only four in ten (39%) knew that about half of all new HIV infections in the U.S. occur among African Americans, and a similar share (40%) knew that half of all new infections in the U.S. occur among people under age 25. Thus, it appears that Americans know more about the HIV/AIDS situation in Africa than they do about their own country.

The American publics' understanding of the AIDS pandemic in Sub-Saharan Africa is dominated by media outlets which, unfortunately, often portray negative imagery and misguided generalizations. For instance, in this same Kaiser Family Foundation survey, when Americans were asked where they get information about HIV/AIDS, a large majority (71%) say that most of the information comes from the media. About half (51%) report they have seen a lot about the problem of AIDS in Africa in the last year, while less state they have seen a lot about the problem of AIDS in the U.S. (34%), Asia (11%), Latin America (7%), and Eastern Europe (5%). And while most Americans are aware of the problem of AIDS in Africa, it is difficult to contextualize this information, because Africa is often depicted in monolithic terms, even though it is a continent of 11.5 million square miles and 53 culturally, linguistically, and religiously diverse nations (Geshekter, 1995).

Distorted media accounts and a general lack of rounded understanding of African nations and culture conspire to maintain the historical, economic and psychological relationship between "us and them" (Clark, 2004; Mayhew, 2002). One such relationship, the association between AIDS and Africa, has become almost reflexive (Geshekter, 1995) with nearly every report beginning with the recitation of the HIV/AIDS statistics for women—Women comprise about half of all people living with HIV worldwide. In Sub-Saharan Africa, women make up 57% of people living with HIV, and three quarters of young people infected on the continent are young women aged 15-24 (UNAIDS/WHO, 2004).

Within this milieu, the ABC approach has been adopted as a public health campaign to combat HIV/AIDS in Africa. However, the ABC health campaign may be unrealistic for many African women and men. In what follows, we provide a feminist reading of the ABC approach, identifying why this health campaign may be problematic. We organize our analysis around each component—abstinence, be faithful, and condom use.

Abstinence

The feminization of AIDS focuses efforts on protecting "vulnerable" women and their children. Female sexuality is constructed around purity, selfrestraint, and the denial of sexual pleasure, with chastity and morality as the underlying logics (Cheng, 2005). Thus, the health message for women is to abstain from premarital and extramarital sex—"Abstinence is the only sure way to prevent sexual transmission of AIDS and other sexually transmitted diseases."

While the health message is clinically accurate, we find moral judgments about self control and sexuality embedded in the call for abstinence. Personal choice over when and how to engage in safer sexual activities is less prominent. And while moral assumptions may not reflect the cultural practices and beliefs of many African women and men, all HIV/AIDS prevention programs funded by the U.S. federal government are required to promote abstinence. Moreover, in the 2004 State of the Union Address, President Bush called for a new emphasis on abstinence-only education, and doubling the funding for abstinence-only programs (Office of National AIDS Policy, 2005). Thus, the U.S. government is not a neutral, philanthropic provider of aid**.** Rather, donations are subject to U.S. political interests that influence policy decisions to support programs and services, which may in fact be incompatible with local needs.

It is not just the content of the abstinence health message, but who has the power to determine the content of the health messages. "ABC has become little more than an excuse and justification to promote their

[U.S. government's] long-standing agenda regarding people's sexual behavior and the kind of sex education they should receive - A for unmarried people, bolstered by advocacy of B, but for most people, anything but C" (Cohen, 2003).

However, abstinence is difficult in practice. Virgins are prized by older men attempting to avoid contacting HIV infections (Gupta, 2000). This places young girls at risk for becoming infected by men who may already be HIV positive. Women also suffer from forced sex. The forces may be physical such as in rape, but women may also be forced into sex purely for survival. In areas which have been badly hurt by economic recession, many women and girls find economic refuge in "sexual networking," or exchanging sexual favors for money, gifts, and protection.

Be Faithful

The "be faithful" message privileges mutually faithful monogamous relationships in the context of marriage as the expected standard of human sexual activity (Collins, Alagiri, & Summers, 2002). However, mounting evidence suggests that married monogamous women are among the groups at greatest risk of infection. For instance, in Kenya and Zambia, data reveal higher rates of infection among young married women (age 15 to 19) than among their sexually active, unmarried (female) peers. These studies found that the rate of HIV infections in husbands was higher than in the boyfriends of sexually active single teenage women. Women in marital relationships were also more frequently exposed to unprotected sex (UNAIDS/WHO, 2004). Women with no economic independence feel constrained to adopt whatever behavior is necessary to protect their marital status, including overlooking their partner's infidelities (Gupta, 2000).

Patriarchal Power is one factor which helps to explain elevated risks for monogamous, married women. Patriarchal Power is often expressed and reproduced through societal norms regarding female sexuality (not abstain from premarital and extramarital sex), and expectations about if and when women are to be married. Patriarchal Power inscribed in the law requires that married women give up their name, property, and personhood. Patriarchal Power is sometimes used to construct the argument that African men are largely unfaithful to their significant others, unwilling to curb their sexual drive, and averse to condom use (Roberts, 2003). Wives are expected to be sexually passive, while husbands are expected to actively pursue sex. Women cannot turn down sex from their husband, and they cannot fully enjoy sex. Thus, wives are dissuaded from participating fully and actively in sex. This may promote the spread of the virus, because some men suggest that their wives "just lie there," and use this as a rationale for soliciting prostitutes (Schoofs, 1999).

Hence, husbands are seen as promiscuous and less deserving of sympathy, while wives are seen as vulnerable and sexually compliant. Sexual compliance is part of the social role and expectations for married women, and is just one example of the way in which Patriarchal Power is intimately intertwined with social justice and HIV prevention. A faithful woman will find it difficult to reduce her risk of contracting HIV/AIDS in an otherwise misogynist material context (Gavey, 1992).

HIV testing is integral to the "be faithful" message, because it lets both partners know their status. However, introducing voluntary testing and counseling would not be easy, as there was a still a lot of fear and stigma surrounding the disease. "Somebody just can't go for testing because it means one thing: you are going find out whether you are going to die or not," (UN Office for the Coordination of Humanitarian Affairs, 2005). There are few positive messages about living with HIV. Therefore, this is still seen as a disease that kills, which makes it extremely difficult to encourage people to be open about their HIV status.

Condoms

Consistent and correct use of condoms is the third component of the ABC health campaign. This message is generally targeted at heterosexual women, and suggests that women should act assertively to control the course of their sexual encounters to ensure that the male partner uses a condom (Gavey, McPhillips, & Doherty, 2001). This message may, however, be problematic for several reasons. First, the discourse of condom use is couched in Western notions of individualism and personal responsibility. This creates contradictions in gender roles and personal identity as a women's desire to be a faithful and committed partner. It also contradicts the cultural significations of sexual intercourse as an expression of monogamy, commitment, love, and trust. This demand for condom use also calls for women to enact assertive sexual behaviors that may go against their feminine identity to acquiesce irrespective of her desire not to have unprotected sex. Some women think of sex as something that happens to them, rather than something they choose. Thus, if a woman's desire to acquiesce overrides her desire to be assertive, then condoms provide little support. For instance, a study in Zambia found that fewer than 25% of the women interviewed believed that a married woman could refuse to have sex with her husband, even if he had been demonstrably unfaithful and was infected. Only 11% thought that a woman could ask her husband to use a condom in these circumstances (UNIFEM, 2001).

Second, this message places women in the position of negotiating with her partner, and suggests the possibility of an individual, rational solution in a relation dominated by the male (Giffin, 1998). At the same time, it ignores the constraints on women's ability to control condom use. Women throughout the world have relatively less control over their sexuality or over their partners' sex behavior (Gupta, 2000). In parts of Sub-Saharan Africa, for instance, poverty and lack of alternatives sometime lead adolescent girls to exchange sex for food, school fees, money, and other commodities, often with men who are much older. Her male partner may force her to engage in unprotected sex, which exposes her to the danger of physical force from their male partner. In this way, health campaigns that promote condom use for women reinforce heterosexuality and the dominance of the male sex drive (Gavey et al., 2001). However, a male partner may also refuse to have sex, which would interfere with a women's sexual pleasure, and jeopardize her chances of continuing a consensual sexual encounter.

Third, condoms carry a stigma. Studies on every continent demonstrate that both men and women perceive condoms for use when having sex with "others," not stable partners. Condoms are for "women of the street, not the home." Even sex workers who are scrupulous about using condoms with clients tend to avoid them with boyfriends and husbands. All too often, condom use has become a sign for the level of trust in a relationship, rather than simply a sensible means of protection (Gavey et al., 2001).

DISCUSSION

While the discourse surrounding the ABC approach appropriates a feminist tone in its concern for women, it is in some ways limited in its promotion of legitimate debate on gender relations. According to Kofi Annan, former UN Secretary General, the ABC approach requests individual change without enacting the societal change that would facilitate women's agency. The driving forces for HIV transmission in southern Africa are linked to structural inequities such as poverty, the economic and social dependence of women on men, and a fear of discrimination that prevents people from openly discussing their status. Women are not able to disclose to their partners that they may have been exposed to HIV in case they are vilified, deserted, and left destitute. Society's inequalities also put them at risk through the lack of access to AIDS treatment, coercion by older men, and men having several partners (UN Office for the Coordination of Humanitarian, 2004). This broader sociopolitical context contributes to the AIDS crisis in Africa. Ecological degradation, migratory labor systems, rural poverty, and civil wars are the primary threats to African lives (Geshekter, 1995).

This, however, does not discount the moral imperative for providing reliable information about HIV/AIDS prevention, and equitable access to condoms and life sustaining medications. In targeting different prevention messages to different groups of people (married women remain faithful, unmarried women abstain, and sexually active women outside of marriage should condomize), the ABC approach fails to consider that a woman may need different messages at different stages of life. For example, a young woman who abstains until marriage is likely to want and need condoms in order to be able to plan childbearing and protect herself from HIV. After a few years of marriage, this same woman may want to become pregnant. However, until non-spermicidal microbicides are developed, this woman has no method for safely becoming pregnant while protecting herself from the risk of HIV/AIDS from a husband who may have other sexual partners (Cohen, 2003). Reproductive health is clearly not distinct from HIV/AIDS prevention, because women often are trying to prevent both HIV/AIDS and unplanned pregnancies simultaneously. The lack of reproductive technologies available to women who are trying to conceive may place them at increased risk.

Massive condom distribution and access to antiretroviral drugs are necessary for preventing both unplanned pregnancies and HIV infection. However, the need simply exceeds the resources of many of the world's poorest nations.

"Abstinence cannot happen in an illiterate community where people have sex early. 'Be faithful' is not going to happen, and 'condomise' - we are ready to use them, but we have no access to money. Condoms are not freely available, and at the price of 500 Ugandan shillings (25 cents) for a packet of three, are luxuries many cannot afford ... people can spend that money on condoms or food ... they are not going to spend their food money on a piece of rubber" (UN Office for the Coordination of Humanitarian Affairs, 2005).

Given these challenges, it seems arrogant and ineffective to impose a Western health campaign without careful adaptation to the various economic, social, and cultural forces that exist in the nations of Africa.

REFERENCES

Cheng, S. (2005). Popularising purity: Gender, sexuality and nationalism in HIV/AIDS prevention for South Korean youths. *Asia Pacific Viewpoint, 46*(1), 7–20.

Clark, D. (2004). The production of a contemporary famine image: The image economy, indigenous photographers and the case of Mekanic Philipos. *Journal of International Development, 16,* 693–704.

Cohen, S. (2003). *Beyond slogans: Lessons from Uganda's ABC experience*. The Guttmacher Report on Public Policy. Retrieved February 8, 2008, from http://www.agi-usa.org/pubs/ib2004no2.html

Collins, C., Alagiri, P., & Summers, T. (2002). *Abstinence only vs. comprehensive sex education: What are the arguments? What is the evidence?* Retrieved February 8, 2008, from http://ari.ucsf.edu/pdf/abstinence.pdf

Gavey, N. (1992). Technologies and effects of heterosexual coercion. *Feminism and Psychology, 2*(3), 325–335.

Gavey, N., McPhillips, K., & Doherty, M. (2001). If it's not on, it's not on: Or is it? Discursive constraints on women's condom use. *Gender and Society, 15*(6), 917–934.

Geshekter, C. (1995). Outbreak? AIDS, Africa, and the medicalization of poverty. *Transition, 67*, 4–14.

Giffin, K. (1998). Beyond empowerment: Heterosexuals and prevention of AIDS. *Social Science and Medicine, 46*(2), 151-156.

Gupta, G. R. (2000). *Gender, sexuality, and HIV/AIDS: The what, the why, and the how*. Retrieved February 8, 2008, from http://www.icrw.org/docs/Durban_HIVAIDS_speech700.pdf

Mayhew, S. (2002). Donor dealings: The impact of international donor aid on sexual and reproductive health services. *International Family Planning Perspectives*, 28(4), 220–224.

Office of National AIDS Policy. (2005). *President Bush discusses HIV/AIDS initiatives in Philadelphia.* Retrieved February 8, 2008, from http://www.whitehouse.gov/infocus/HIVAIDS/

Office of the United States Global AIDS Coordinator. (2005). *Engendering bold leadership: The President's emergency plan for AIDS relief*. First Annual Report to Congress. Retrieved February 8, 2008, from http://www.state.gov/documents/organization/43885.pdf

Roberts, C. (2003). *History repeats itself in global AIDS struggle.* ifeminist.com. Retrieved February 8, 2008, from http://www.ifeminists.net/introduction/editorials/2003/0415roberts.html

Schoofs, M. (1999). AIDS: The agony of Africa: Part 5: Death and the second sex. *Village Voice*. Retrieved February 8, 2008, from http://www.villagevoice.com/news/9948.schoofs.10565.1.html

Sontag, S. (1990). *Illness as metaphor and AIDS and its metaphors.* New York: Farrar Straus and Giroux.

Treichler, P. (1999). *How to have theory in an epidemic: Cultural chronicles and AIDS*. Durham, DC: Duke University Press.

UN Integrated Regional Information Networks. (2003). *Sudan: HIV/AIDS not as easy as ABC*. Retrieved February 8, 2008, from http://www.aegis.com/news/IRIN/2003/IR031023.html

UNAIDS/WHO. (2004). *AIDS epidemic update*. Retrieved February 8, 2008, from http://www.unAIDS.org/wad2004/EPI_1204_pdf_en/EpiUpdate04_en.pdf

UNIFEM. (2001). *Fact sheet on gender and HIV/AIDS*. Retrieved February 8, 2008, from http://www.unifem.org/index.php?f_page_pid=30

Waldby, C. (1996). *AIDS and the body politic.* London: Routlege.

KEY TERMS

ABC Health Campaign: The centerpiece of the prevention component of President Bush's Emergency Plan for AIDS Relief. Abstinence should take precedence for people who are not in a relationship. Those who are in a relationship should remain faithful to their partners. And if the first two strategies fail for any reason, condoms should be used to prevent the transmission of HIV.

Acquired Immune Deficiency Syndrome (AIDS): AIDS is a human disease characterized by progressive destruction of the body's immune system. It is widely accepted that AIDS results from infection with HIV. Although treatments for both AIDS and HIV exist, there is no known cure.

Afropessimism: An unremittingly bleak view of Africa. The persistence of the menacing image of Africa in the West is highlighted at the present time by the AIDS pandemic.

Human Immunodeficiency Virus (HIV): Immunodeficiency means having a faulty immune system so that a person can become very ill or die from a disease that others can fight off. Acquired means that HIV is passed from person to person through blood or other bodily fluids, through a transfusion of infected blood, to a baby from its mother, through use of contaminated hypodermic needles, or through sexual contact with a person who has the disease.

Sub-Saharan Africa: The region of Africa south of the Sahara desert. Sub-Saharan Africa, one of the poorest regions in the world, still suffering from the legacies of colonial conquest and occupation, neocolonialism, and internal conflict. The region is comprised of 48 nations, many of which are among the least developed countries in the world

Acoustic Feature Analysis for Hypernasality Detection in Children

Genaro Daza
Universidad Nacional de Colombia Sede Manizales, Colombia

Luis Gonzalo Sánchez
Universidad Nacional de Colombia Sede Manizales, Colombia

Franklin A. Sepúlveda
Universidad Nacional de Colombia Sede Manizales, Colombia

Germán Castellanos D.
Universidad Nacional de Colombia Sede Manizales, Colombia

INTRODUCTION

In the treatment of children with fixed Cleft Lip and Palate (CLP), problems such as hipponasality and hypernasality, which are related to vocal emission and resonance, might appear. Nevertheless, according to the report presented in Castellanos et al. (2006), hypernasality is more frequently found than hipponasality (90% vs. 10%). The interest shown in hypernasality detection is that its occurrence points out problems of anatomical and neurological sort that are also related to the peripheral nervous system (Cairns, Hansen, & Riski, 1996). The presence of hypernasality, understood as the leak of nasal air and compensatory articulations, leads to low intelligibility of speech. This declining of the subject's communication capabilities may end up in behavioral and social changes.

In velopharyngeal learning, the distortion of the acoustic production leads up to nasalized voice. Moreover, since air loss or nasal leak is massive, articulator mechanisms are compromised. The patient can not speak clearly and intelligibly, and thence they replace their velum palatine sphincter by glottal articulation that allows for clearer speech: /p/, /t/, /k/, /b/, /d/, /g/ come from glottal stops, while sounds like /ch/, /s/, /t/, /j / are accompanied by hoarseness (Habbaby, 2002). Although hard palate has been repaired surgically, it might not provide velopharyngeal competence necessary for a normal speech production. Even if the palate is potentially capable after surgery, previous speech habits might have developed compensatory articulations or physiologic compensations that aimed to approx to intelligibility that enhance the number of pathological patterns in speech. As a result, compensatory articulations persist generally, even after undergoing post technical or post surgical manipulation that had forecasted a plenty shutting. Thus, they have to be fixed before increasing the performance of the velopharyngeal sphincter throughout language therapy.

In the last years, there has been a growing interest for acoustic speech analysis (ASA) as an alternative method for diagnosis and treatment in identifying functional disorders in children's voices (González, Cervera, & Miralles, 2002; Niedzielska, 2000; Niedzielska, Glijer, & Niedzielski, 2001). This type of analysis exposes great advantages over traditional methods due to its noninvasive nature and potential to obtain a quantitative measure on the clinic state of the larynx and vocal tract.

Acoustic features or objective parameters are frequently used to represent the pathologic voice on held vowels (Hadjitodorov & P.Mitev, 2002; Kent, Vorperian, & Kent, 200; Yu, Ouaknine, Revis, & Giovanni, 2001). However, such vectors are limited in their robustness because of their estimation complexity in real conditions with perturbations of nonstationary structure (Gupta & Gilbert, 2001). Although several analysis of effectiveness have been made around the different kinds of proposed features for objective evaluation of speech disorders (Frohlich & Michaelis, 2000; Yu et al., 2001), they can not be taken as a standard set of parameters for hypernasality identification because each disorder affects differently diverse aspects of speech emission.

The present work analyzes the statistical effectiveness of different acoustic features in the automatic identification of hypernasality. Acoustic features reflect part of information contained in perceptual analysis; in part, due to their estimation is derived directly or indirectly from the vocal cords behavior. Consequently, it is convenient to apply multivariate analysis techniques in determining the effectiveness of voice features. The effectiveness is studied by using multivariate analysis techniques that are meant for feature extraction and feature selection, as well (latent variable models, heuristic search algorithms).

BACKGROUND

Nasalization and Nasal Emission

Nasalization is defined as the link between nasal cavity and the rest of the vocal tract; while nasal emission refers to abnormal air loss through nasal route. This abnormal leakage reduces intra-oral pressure causing distortion in consonants. When air loss turns into an audible reblowing, the nasal emission is more obstructive and speech is seriously affected. Nasality commonly named hypernasality refers to low speech quality, which results from inappropriate adding of the resonance system to vocal tract. Conversely to nasal emission, nasality does not involve large flows of nasal air, so that there is no significant change in intra-oral air pressure. For this pathology, identification studies based on signal modeling (specialized diagnosis) can be related to acoustic features by using pattern recognition techniques.

Nasalization Model

Considering that normal voice is made of resonances at different frequency formants F_k, the following acoustic model (Cairns et al., 1996) is proposed:

$$S_n(\omega) = \sum_{k=1}^{K} F_k(\omega)$$

In contrast to normal voice, nasalized voice is the appearance of the antiformants $\widehat{F}'$ and nasal formants $\widehat{F}$:

$$S_h(\omega) = \sum_{k=1}^{K} F_k(\omega) - \sum_{l=1}^{L} \widehat{F}'_k(\omega) + \sum_{m=1}^{M} \widehat{F}_m(\omega)$$

It has been suggested that the intensity in the reduction of the first formant is a primary indicator of nasality. In Cairns et al. (1996), the superior formants are filtered in such a way that filtered normal voices will have one component, while nasalized voices will correspond to a signal with several components, which can be estimated using Teager's instant power operator.

In a general way, patients with CLP manifest nasality with a deficiency of the velopharyngeal port, coupling their nasal cavity to vocalic sounds that gives place to appearance of an additional resonance in the amplitude-frequency feature of the vocal tract, and so a noticeable decrease in the formants F1 and F2 (Baken, 1996) as shown in sections (a) and (b) of Figure 1.

ACOUSTIC FEATURES AND MULTIVARIATE ANALYSIS

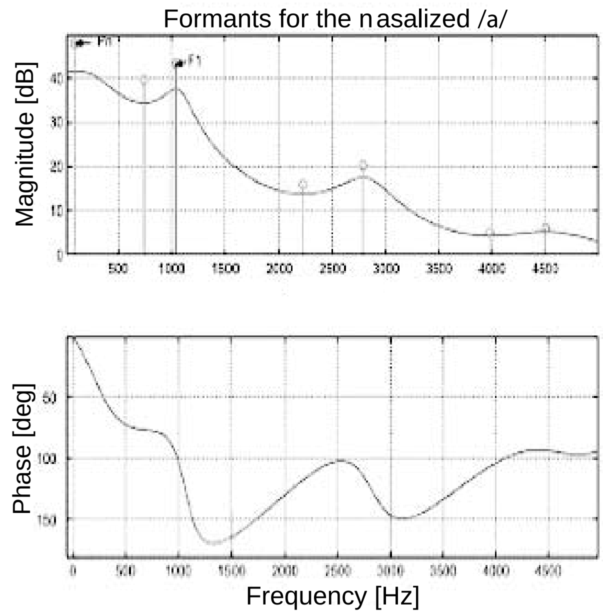
Acoustic features can be split into two categories according to the acoustic properties to be measured. Based in additive noise, among them: Harmonic to Noise Ratio (HNR), Normalized Noise Energy (NNE), Glottal Noise Excitation (GNE), defined as the noise estimation and it is based in the assumption that resulting glottal pulses from collisions of vocal folds head to a synchronous excitation of the different band frequencies, Normalized Error prediction (NEP) that can be expressed as the relationship between geometric and arithmetic means of spectral model, and Turbulence Noise Index (TNI).

Other acoustic features are associated to frequency modulation noise, among them pitch or fundamental period of the signal and jitter, which is defined as the average variation percentage between two consecutive values of pitch. In addition, there are considered features associated to parametric models of speech generation. Among them: cepstral coefficients derived from linear prediction analysis (LPC Linear Prediction Coefficients), cepstral coefficients over pounded frequencies scale (MFCC Mel-Frequency Cepstrum Coefficients), and RASTA coefficients (Relative Spectral Transform) (Castellanos, Castrillón, & Guijarro, 2004).

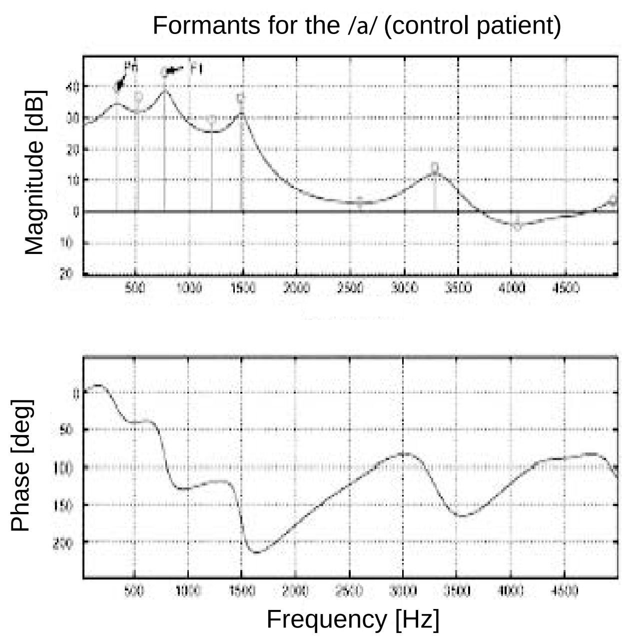
Multivariate Analysis

Latent Variable Models. The set X of measured variables can be seen as the linear combination of generative factors Z and perturbations ε over the measurement process.

Figure 1. Pitch energy for distinguishing glottal stop



(a) Pitch Energy (Pathological case)



(b) Pitch Energy (Normal case)

$$X = WZ + \varepsilon$$

Where W is the weight matrix that assigns the influence of each factor on each observed variable.

Different assumptions about how this factors and perturbations distribute derive in the well known feature extraction techniques. In particular, we consider Principal Component Analysis (PCA), Factor Analysis (FA), and Independent Component Analysis (ICA), for our experiments. A detailed description of these topics can be found in Hyvärinen, Karhunen, and Oja (2001), Jolliffe (2002), and Rencher (2002).

Feature Selection. The proposed methodology for reducing the initial spaces is based on strong relevance analysis, which studies the correlation among features. Reduction methodology considers combining heuristic techniques such as Sequential Forward Search (SFS) (Webb, 2002) with multivariate analysis for further evaluation of feature subsets. Particularly, multivariate analysis of variance (MANOVA) is used as the intrinsic cost function to assess the discriminative power of a given set (Rencher, 2002).

EXPERIMENTAL BACKGROUND

Database

The sample is constituted by 68 children. Classes are balanced (34 patients with normal voice and 34 with hypernasality) and evaluated by specialists. Each recording is conformed by five words of Spanish language: /coco/, /gato/, /jugo/, /mano/, and /papá/. Signals are acquired under low noise conditions using a dynamic, unidirectional microphone (cardioide). Signal range is between (−1, 1).

Table 1. Variable elimination by preprocessing

Word of Analysis	Reduced Space Dimensionality (Reduced/Original)
/coco/	31/40
/gato/	29/40
/jugo/	12/40
/mano/	16/20
/papá/	28/40

INITIAL FEATURE SPACE

A complete set of considered features are: Pitch (F0) (mean value and standard deviation) (Manfredi, D'Aniello, Bruscaglioni, & Ismaelli, 2000; Sepúlveda, Castellanos, & Quintero, 2004), Jitter, Shimmer (Childers, 2000), TNI (Hadjitodorov & P.Mitev, 2002), HNR (Yumoto & Gould, 1982), NNE (Kasuya, Ogawa, Mashima, & Ebihara, 1988), and Mel Frequency Cepstrum Coefficients (MFCC) (13 coefficients) (Huang, Acero, & Hon, 2001). Depending on the analysis word, the aforementioned features can be extracted for one or two voiced segments, which imply that 20 or 40 features represent each register.

Data Preprocessing

The main target of data processing is to reduce or even eliminate the influence of measurement errors. Among them: systematic errors during acquisition, occasional failures in the measurement instruments, and so on. It is also used for controlling the homogeneity principle of the different statistical properties of phenomena in analysis. Data preprocessing consists of analyzing odd logs for each feature and assessing their normality (Peña, 2002).

Classification

The employed classifier is Bayesian. Five classifiers of this kind are used; each of them for analyzing the error between classes (hyper-nasal and normal or control) for each one of the words previously mentioned. Besides, a leave one out cross-validation was conducted to observe the variation in the classifier's parameters and its generalization capability (Webb, 2002).

RESULTS

Acoustic feature effectiveness can be measured according to classification performance. Next, the results are displayed for each one of the stages. The abnormal value detection is conducted for each feature. It allows making clear the quality of the measurements. Those features with 10% or more outliers in either population samples are discarded. Acoustic features results are shown on Table 1. Average reduction for word-set is between 30% to 35%.

Table 2. Classification accuracy of unprocessed data

	Methods							
	PCA		FA		ICA		MANOVA	
Word	**N.L.V.**	**C.A.**	**N.L.V.**	**C.A.**	**N.L.V.**	**C.A.**	**N.L.V.**	**C.A.**
/coco/	12	0,7059	18	0,6912	10	0,6176	11	0,8382
/gato/	10	0,9557	5	0,9265	10	0,9559	3	0,9559
/jugo/	5	0,9853	5	0,9412	5	0,9853	1	0,8824
/mano/	4	0,5441	10	0,75	4	0,5441	4	0,8235
/papá/	9	0,7206	17	0,7206	9	0,7206	7	0,8088

Table 3. Classification accuracy of preprocessed data

	Methods							
	PCA		FA		ICA		MANOVA	
Word	**N.L.V.**	**C.A.**	**N.L.V.**	**C.A.**	**N.L.V.**	**C.A.**	**N.L.V.**	**C.A.**
/coco/	7	0,5882	15	0,7353	3	0,6324	7	0,8088
/gato/	9	0,9706	6	0,9853	7	0,7353	3	0,9265
/jugo/	3	0,9412	6	0,9559	3	0,9412	1	0,8824
/mano/	4	0,5882	8	0,7206	4	0,5882	4	0,7647
/papá/	8	0,7059	13	0,75	5	0,6324	6	0,8088

In what concerns normality verification, Gaussian structure of data can be achieved by using a hypothesis test; in this case, a kurtosis test.

The paramount purpose of effective variable selection and feature extraction is to find a reduced representation set that allows for better discrimination among classes. As previously mentioned, feature extraction is carried out by using linear models of latent variables and feature selection is accomplished by using a heuristic technique for growing a subset called "filter." MANOVA becomes then the cost function. The effective feature selection is applied twice: (a) without data preprocessing (Table 2), and (b) after data preprocessing (Table 3). Although, classification accuracy decreases after preprocessing, this is due to size of samples that are small to provide better estimates. Overall results displayed in Table 4 show the MFCCs as good descriptors for detecting hypernasality (N.L.V. stands for number of latent variables or selected features and C.A. for classification accuracy).

FUTURE TRENDS

In the case of hypernasality detection, it would better to design specialized diagnostic features that reflect the nature of the irregularities in the functional state of speech. Such features might offer similar detection levels compared to acoustic features, nonetheless being less computationally complex. Moreover, training a classifier would not be necessary; the compromise degrees of resonance in the pathology could be estimated directly from the estimated features.

Detecting the hypernasality degrees is important because in diagnosis as well as in treatment and surgery, the specialists need to have available a taxonomic reference system that can be useful for relating the different compromise degrees of resonance to velopharyngeal patterns, generating, then, different alternatives for treatment while excluding others.

A

Table 4. Selected variables

Word of analysis	Selected Variables
/coco/	TNI, HNR and MFCCs.
/gato/	MFCCs.
/jugo/	MFCCs.
/mano/	MFCCs.
/papá/	MFCCs.

CONCLUSION

Acoustic features can be applied to discriminate pathological from normal speech with fair accuracy. One of the main problems for clustering voice features in hypernasality analysis is their sensibility to acoustic properties. This mutual dependency might be one of the reasons for their contradictory interpretation along literature. Consequently, phonemes of velar articulation patterns with enough discriminating power ought to be identified (lip posture detection) and vocal emission (acoustic analysis). Previous e-tudies have suggested the use of MFCCs in pathological voice detection (Godino-Llorente, Gómez-Vilda, & Blanco-Velasco, 2006). This situation has been corroborated here, since MFCCs were the resulting feature after applying variable selection algorithms.

REFERENCES

Baken, R. J. (1996). *Clinical measurement of speech and voice*. Thomson Delmar Learning.

Cairns, D. A., Hansen, J. H. L., & Riski, J. E. (1996). A noninvasive technique for detecting hipernasal speech using a nonlinear operator. *IEEE Transactions on Biomedical Engineering.*

Castellanos, G., Castrillón, O., & Guijarro, E. (2004) Multivariate analysis techniques for effective feature selection in voice pathologies. In *XXII Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB 2004),* Santiago de Compostela.

Castellanos, G., Daza, G., Sánchez, L., Castrillón, O., & Suárez, J. (2006). Acoustic speech analysis for hypernasality detection in children. In *IEEE EMBC.* New York.

Childers, D. G. (2000). *Speech processing and synthesis toolboxes*. John Wiley and Sons, Inc.

Frohlich, M., & Michaelis, D. (2000). Acoustic voice analysis by mean of the hoarseness diagram. *Journal of speech, language and hearing research, 3*(43).

Godino-Llorente, J. I., Gómez-Vilda, P., & Blanco-Velasco, M. (2006). Dimensionality reduction of a pathological voice quality assessment system based on gaussian mixture models and short term cepstral parameters. *IEEE Transactions on Biomedical Engineering, 53*(10), 1943-1953.

González, J., Cervera, T., & Miralles, J. L. (2002). Análisis acústico de la voz: Fiabilidad de un conjunto de parámetros multidimensionales. *Acta Otorrinolaringol*.

Gupta, M., & Gilbert, A. (2001). Robust speech recognition using wavelet coefficient features. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding* (pp. 445- 448).

Habbaby, A. (2002). *Enfoque integral del niño con fisura labiopalatina*. Buenos Aires: Médica Panamericana.

Hadjitodorov, S., & P.Mitev. (2002). A computer system for acoustic analysis of pathological voices and laryngeal diseases screening. *Medical Engineering and Physics*.

Huang, X., Acero, A., & Hon, H.-W. (2001). *Spoken language processing*. Upper Saddle River, NJ: Prentice Hall.

Hyvärinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis.* In. S. Haykin (Ed.). John Wiley and Sons, Inc.

Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). Springer.

Kasuya, H., Ogawa, S., Mashima, K., & Ebihara, S. (1988, November). Normalized noise energy as an acoustic measure to evaluate pathologic voice. *Journal of the Acoustical Society of America, 80*(5).

Kent, R., Vorperian, H., & Kent, J. (2003).Voice dysfunction in dysarthria: Application of the multidimensional voice program. *Journal of communication Disorders*, *36*(4), 281-306.

Manfredi, C., D'Aniello, M., Bruscaglioni, P., & Ismaelli, A. (2000). A comparative analysis of fundamental frequency estimation methods with application to pathological voices. *Elsevier Medical Engineering and Physics, 22.*

Niedzielska, G. (2000). Acoustic analysis in the diagnosis of voice disorders in children. *International Journal in Pediatric Otorhinolaryngology*.

Niedzielska, G. (2001). Acoustic analysis in the diagnosis of voice disorders in children. *International Journal of Pediatric Otorhinolaryngology*, *57*(3), 189-193.

Peña, D. (2002). *Análisis de datos multivariantes*. C. F. Madrid (Ed.). McGraw Hill.

Rencher, A. C. (2002). *Methods of multivariate analysis* (2nd ed.). John Wiley and Sons, Inc.

Sepúlveda, F. A., Castellanos, G., & Quintero, C. (2004). Estimación robusta del pitch usando transformada wavelet. In *III congreso latinoamericano de ingeniería biomédica III CLAEB*, Joao-Pessoa, Brazil.

Webb, A. R. (2002). *Statistical pattern recognition* (2nd ed.). John Willey and Sons.

Yu, P., Ouaknine, M., Revis, J., & Giovanni, A. (2001). Objective voice analysis for dysphonic patients: A multiparametric protocol including acoustic and aerodynamic measurements. *Journal of Voice*, 15(4), 529-542.

Yumoto, E., & Gould, W. J. (1982, June). Harmonics-to-noise ratio as an index of the degree of hoarseness. *Journal of the Acoustical Society of America, 71*(6).

KEY TERMS

Acoustic Features: Physical characteristics of speech sounds such as color, loudness, amplitude, frequency and so forth.

Bayesian Classifier: A simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions.

Heuristic Search: Any search strategy which makes use of heuristics to suggest the best nodes to consider at each stage of the search.

Hypernasality: A quality of voice in which the emission of air through the nose is excessive due to velopharyngeal insufficiency; it causes deterioration of intelligibility of speech.

Latent Variable: A variable in the model that is not measured.

Multivariate Analysis: An analysis of the relationships between more than two variables.

Nasalize: To make nasal or produce nasal sounds.

Adoption of ICT in an Australian Rural Division of General Practice

Patricia Deering
Victoria University, Australia

Arthur Tatnall
Victoria University, Australia

INTRODUCTION

Many information technology (IT) products have been developed to support medical general practitioners (GPs) in all aspects of their work (GPSRG, 1998), and much research and development in this area has already been done. It is apparent, however, that GPs are not making as much use of these systems as they could. Our research showed that there is still reluctance, in particular from many rural general practitioners to fully implement information and communication technologies (ICT) in primary health care in rural Australia (Everitt & Tatnall, 2003). While a simple analysis of the statistics of the numbers of computers in medical practice shows that there are computers in most general practices, it is not so clear how, or even whether, they are being used. Rural GPs, however, operate very much in the mode of small business (Burgess & Trethowan, 2002). Some national research shows that GPs use ICT mainly for administrative and some clinical functions, but that much less use is made of online functions (NHIMAC, 1999; GPCG, 2001). This is even more pronounced for rural GPs.

It is clear that the introduction of the Practice Incentives Program by the Australian Government in 1998 dramatically accelerated the adoption of computer technologies (Tatnall, Everitt, Wenn, Burgess, Sellitto, & Darbyshire, 2004); however, the question still remains—do the GPs use the computers they have acquired, and if so, for what? If they do use computers, then why do some GPs use them more extensively than others? The answers to these questions are not simple and do not appear to be related to the age of the GPs. The aim of the research project described in this chapter was to explain in detail the reasons for adoption or nonadoption of ICT in these medical practices and to develop a model that illustrates the process of decision-making in relation to ICT innovations.

The view that research such as this needs to be grounded in the real world of health practice is supported by continuing increases in health costs and complexity of health delivery. The amount of medical information available to medical practitioners has increased the need to find better ways to manage medical practice, and although hospitals are also experiencing these problems, it is on general practitioners that much of the information management burden falls. Many GPs are examining ways that information can be better managed, and various types of information management systems are becoming an important focus of their work.

USE OF ICT BY MEDICAL GENERAL PRACTITIONERS

While one might expect that highly educated professionals such as most GPs would be at the forefront of the information management revolution; however, our research has shown that this is not entirely the case (Everitt & Tatnall, 2003).

It appears that the slow uptake has continued to some degree in all areas of medical general practice, despite continued support and promotion of computer use (Tatnall et al., 2004). The Commonwealth Department of Health and Aged Care, along with the General Practice Computing Group, also report that general practitioners in Australia are still being encouraged via their Divisions of General Practice to adopt electronic information systems to enhance clinical and practice management. However, Richards (1999) notes, "The adoption of computers by Australian general practitioners has been slow in comparison with other English speaking countries." It is clear that over a long period of time, much research and development have been done on the use of medical ICT, and many products

Table 1. General practice computer use (GPSRG, 1998)

Use of Computers	Level of Usage (%)
Administrative functions	85
Clinical functions	76
Script writing	60
Generating referral letters	57
Receiving results electronically and running recall systems	57

have been developed for GPs. Despite this, however, reluctance on behalf of GPs to use ICT continues and appears to be a fact of life for many of the rural members of the profession.

Table 1 represents the main uses for computers that were identified by a 2001 General Practice Computing Group (GPCG) study of 1,202 GPs.

METHODOLOGICAL FRAMEWORK FOR THE RESEARCH

Many factors and entities are involved in determining how GPs adopt and use ICT, and any approach that ignores the inherent complexity of this sociotechnical situation is unlikely to produce useful answers. This chapter will argue that a qualitative approach using the Actor-Network Theory (ANT) allows for the views of all subjects to be fully documented and explained. This approach was necessary because the complexity of health delivery in Australia has increased the need to manage information in medical practices, leading to a multistakeholder environment involving human and nonhuman entities. Our research aimed to explain the patterns of computer use by rural GPs in an Australian Division of General Practice, and to draw out the factors that contribute to patterns of computer use. In undertaking this research, we compared two theoretical approaches that seek to explain the adoption and use of computers by GPs: Innovation Translation (from the Actor-Network Theory) and the better-known theory of Innovation Diffusion.

In regard to the theoretical approach being taken, there has been a search for a seamless design approach to research in ICT for several decades, but often the methodology for ICT research has been transferred from science. This has not always been successful, as these approaches imply that the scientific method produces reliable knowledge because consistency, dependability, and regularity are the benchmarks for science. However, according to Latour (1987b), despite the fact the scientists aim to subscribe to morally and technically efficient methods that are impersonal and objective with results that are open to everyone, research seldom happens in this manner because science reacts to social pressures just like any other ordinary and mundane system of action.

In his discussion of the Innovation Diffusion model, Rogers (1995) asserts that an individual's decision to adopt an innovation is not an instantaneous act, but rather a process that occurs over time, consisting of a series of actions and decisions. Anything new causes uncertainty, and so there is a lack of predicability and people will seek information to change that situation. In this model, diffusion could be considered a process of information exchange aimed at reducing uncertainty. "The new ideas upon which an innovation is based are communicated over time, through various types of communication channels, among members of a social system through which the innovation diffuses" (Rogers, 1995). There are thus four main elements of any theory of innovation diffusion: characteristic of the innovation itself; the nature of the communication channels; the passage of time; and the social system. Diffusion is considered to be an information exchange process among members of a communicating social network particularly concerned with the characteristics of the innovation. A simplistic view of this project might be that the GPs will make their adoption decisions primarily because of the characteristics of ICT, and would miss other influences due to interbusiness interactions and the backgrounds of the people involved.

While diffusion models (Rogers, 1995) of innovation have had considerable success in explaining the movement (diffusion) of technology in the large scale, they have had much less success in explaining *individual* cases of technology adoption or nonadoption (Latour, 1986; Tatnall & Gilding, 1999). Although Rogers' (1995) approach has increasingly moved toward the inclusion of social factors to explain the diffusion and adoption of technology, we argue that it has not gone far enough. In the field of technology studies, there is an increasing body of work that suggests that there are many complex social factors involved in the interaction of society and technology (Good 1990; Latour, 1987a, 1988; Mol & Law, 1994) and that any process of technology adoption in such an environment must inevitably involve a set of complex negotiations among all those involved. Furthermore, the final outcome typically will involve any devised product being *translated* (Law, 1992) from the form initially proposed to a form suitable for actual use. Most GPs operate in a small business environment, and research has suggested (Tatnall, 2002; Tatnall & Lepa, 2003) that adoption of technology in small business and by special groups can be better explained using an approach such as *innovation translation* (Callon, 1986a; Latour 1986, 1996), for which Latour (1992) claims several advantages: respect to the actors and their individual differences; allows for all entities; pays respect to the outcomes; failures have the same explanations as successes so no hierarchy of dominance is produced; all links in the network are accounted for; and when information is translated from one form to another, it can still be credited to its origins.

The translation model (Latour, 1996) posits the continuous *transformation* of an innovation into new forms by all those who touch it. The essentialist tenets of innovation diffusion—that there is some property in the innovation, the society, or the potential adopter that facilitates diffusion—differ radically from the translation view in which it is the potential adopters who hold the key in their actions. There are occasions when diffusion does not occur despite the excellence of the idea or the technical quality of the innovation, and the diffusion model finds these difficult to explain. Innovation translation concentrates on issues of network formation and investigates the human and nonhuman alliances and networks built up by the various actors involved. It concentrates on the negotiations that allow the network to be configured by the enrolment of both human and nonhuman allies, and considers the

Table 2. Innovation diffusion and innovation translation (McMaster, Vidgen, & Wastell, 1997)

	Innovation Diffusion	**Innovation Translation**
Discipline of origin	Social science and communication theory	Semiotics, sociology of scientific knowledge
Definition of innovation	A technology perceived to be new by a potential adopter	A technology that has yet to be black-boxed
Shaping of innovation	Generally decided at design stage	Shaped by the network of actors
Communication	Communication channels categorized as cosmopolite vs. localite, mass media vs. interpersonal	Translations are made by actors in enrolling the innovation
Time	Speed of decision to innovate, rate of adoption is important	Network dynamics in enrollment, control and dissemination are what matter
The social system	Sharing of interests of human actors	Interessement among actors (human and nonhuman) and goals
Spread of ideas	Diffusion from a single source	Occurs through social/network interactions
Origin of power	Usually vested in a single powerful entity or change agent	The result of interactions among entities in the network
Motivating force	Early adopters provide the example for others to follow	Resides in the network of actors—habits, routines, and institutional behaviors inhibit movement
The technology	Changes made to form and content of the technology as a result of experience during implementation (re-invention)	Technology is translated through being enrolled, regardless of whether its form or content is modified
Socio-technical stance	Social system and technology are separate, diffusion is the adoption of technology by a social system	The social system and the technology are inseparable

IT system's characteristics only as network effects resulting from association (Wenn, Tatnall, Sellitto, Darbyshire, & Burgess, 2002). Innovation translation theory suggests that it is not any *innate properties* of these systems that are important, but rather *network associations* (Callon, 1986b; Bijker, Hughes, & Pinch, 1987), such as the extent to which the medical practice has been enrolled in the advantages of the new system. It looks at the process of redefinition in which the medical practice tries to seek compromises from the IT system and how the system imposes definitions of how the technology should be used; how it "interested" the medical practice and then got them to follow its interests, so becoming indispensable to them. What is finally adopted for this task is not the IT system originally examined as such, but rather a translation of this in which it becomes a tool for their particular medical practice use. Innovation translation is thus more attuned to the human and political issues involved in organizational decision-making, and so offers a useful approach to modeling innovation in small business (Tatnall, 2002).

Innovation diffusion and innovation translation are based on very different philosophies. Table 2 illustrates some of the main differences between these two innovation models.

RESULTS AND CASE STUDY

A significant difficulty experienced in collecting data for this project related to making contact and establishing working relationships with individual GPs, who typically must be accessed via their practice manager. As GPs have such a busy schedule, making time to take part in a research project can be a big time sacrifice and, as they are aware of this, the practice managers are often reluctant to facilitate access. We have learned to regard the practice managers as some of the key contacts in the research project, and as important people who were not at first acknowledged as being important to the study. They act as gatekeepers of the information, so to speak, and are important contacts in reviewing how IT works in medical practices. Another issue arose around doing qualitative research in an industry that is only just coming to grips with research of this type and the processes involved. Continually having to justify the premise for qualitative research certainly keeps the researcher on his or her toes and ensures that the source of funding is given credit. The study also provided a chance to sell qualitative practice in the health sector and contribute in a very positive manner to the development of research methodology in this sector.

The following data relate to a rural Division of General Practice not far from Melbourne, Australia, where this study was conducted. The data come from 98 participating GPs (approximately two-thirds of those currently practicing in this division), of whom 89 were male and nine were female. Of these GPs, 42 practiced full-time and 56 part-time. The most common practice size comprised four GPs (there were 10 of these practices) followed by four larger practices (with more than four GPs) and two solo medical practices. There were 55 GPs in the age range 25 to 35, 27 from 36 to 45, and 16 older than 45. Each practice made some use of computers, and all had been doing so for at least 12 months, with about 60% using them for both clinical and administrative purposes and the remainder using them only for practice administration.

In addition to this quantitative data, the following brief case study provides a snapshot of one medical practice of several GPs located in a small country town. The Practice Principle was the sole decision maker and is very active in promoting adoption of ICT and Information Management. This practice reports wide use of computers for generating prescriptions, investigation request slips, current medication lists, and referral purposes, as well as for the usual administrative needs. The principle of this practice discussed the following concerns:

- Changing the way the doctors work
- The high cost of changing and updating computer technologies
- Doctors learning to type
- Losing patients who do not like the GP/patient/computer interface
- Losing information by accident (erasing information)
- Confidentiality
- Losing valued staff as change to computerization take place
- Money loss as productivity is reduced while learning new technologies
- Security of electric power supply in the country

The GPs from this practice expressed the following concerns:

A

- **Technical issues:** The effect of computer use in the patient and doctor relationship
- **The work environment:** Trialing programs before use
- **Training and education:** A forum to discuss ideas about computers

The following comments taken from these interviews show a very positive attitude to computerization in this division and that the quantitative data previously provided shows only half the picture.

- "Don't be afraid, take a step. The only way to go is forward, unless you want to be anxious or depressed."
- "Electronic information is much easier to transfer from system to system than ink on paper. Lock-in isn't the problem it's thought to be, and the software industry has committed to fight it."
- "Plan for a useful life of computer hardware and software of around three years. The tax man says you can depreciate computers at 36%. If your system lasts functional longer than three years, you have earned a bonus."
- "Get a computer for home, connect it to Internet, and use it. There's no excuse, either for the kids, for yourself, or for your partner."
- "Home computers and modems are at commodity prices. Learn in the privacy of your own home."
- "Network with GPs who can share their computing experiences. Knowledge in this field has a very short shelf life. Anything in print with a cover is out of date. Your absolute best resource is the grapevine. The Internet is a superb grapevine."
- "Computerise your records if there is opportunity. Major opportunities include starting a new practice, changing from cards to the RACGP record system, and implementing health summaries. If your apple cart is stable, it may be best to let it be for a little while."

This small snapshot makes use of an actor-network analysis of the data gained through interview, observation, and discussion, and through a short survey. The information sought related to the experiences of the individual GPs rather than to their ability to use computers. As software and hardware issues are not clearly listed in these concerns, it is suggested that there has been an improvement in both since earlier days of computer introduction These concerns might imply that there is still a lot to be achieved by GPs, and that there may be logical steps in the adoption of computers into a practice (if steps are taken in a logical process, some of the concerns may be eliminated). The comments we have included indicate that use of information technologies does not develop in isolation, and similarly, these technologies do not develop free from social influence. The research indicates that the two shape each other, and that GPs and their practice staff will continue to struggle with the pace of change.

CONCLUSION

In this chapter, we have described how, by use of an actor-network theory, we are one step closer to a more detailed understanding of the relationships between information technology and its use in general practice. This choice was motivated by the way the actor-network theory offers a language and vocabulary for describing the many small, technical, and nontechnical mechanisms that go into the building and use of information infrastructures in rural general practice. The actor-network theory goes a long way to describe how actions are enabled and constrained. Information technologies do not develop in isolation free from social influence, and we have discussed how the two shape each other and how GPs and their practice staff continue to struggle with the pace of change.

REFERENCES

Bijker, W.E., Hughes, T.P., & Pinch, T.J. (Eds.). (1987). *The social construction of technological systems: New directions in the sociology and history of technology*. Cambridge, MA: MIT Press.

Burgess, S., & Trethowan, P. (2002). *GPs and their Web sites in Australia: Doctors as small businesses*. Las Vegas: OneWorld.

Callon, M. (1986a). The sociology of an actor-network: The case of the electric vehicle. In M. Callon, J. Law, & A. Rip, *Mapping the dynamics of science and technology* (pp. 19–34). London: Macmillan Press.

Callon, M. (1986b). Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St. Brieuc Bay. In J. Law, *Power, action & belief. A new sociology of knowledge?* (pp. 196–229). London: Routledge & Kegan Paul.

Everitt, P., & Tatnall, A. (2003). *Investigating the adoption and use of information technology by general practitioners in rural Australia and why this is less than it might be*. Perth: ACIS.

Good, B.J. (1990). *Medicine, rationality and experience: An anthropological perspective*. Cambridge, UK: Cambridge University Press.

GPCG. (2001). *Measuring IT use in Australian general practice*. Brisbane: General Practice Computing Group, University of Queensland.

GPSRG. (1998). *Changing the future through partnerships*. Canberra: Commonwealth Department of Health and Family Services, General Practice Strategy Review Group.

Latour, B. (1986). The powers of association. In J. Law, *Power, action and belief. A new sociology of knowledge? Sociological Review monograph 32* (pp. 264–280). London: Routledge & Kegan Paul.

Latour, B. (1987a). *Science in action*. Cambridge, MA: Harvard University Press.

Latour, B. (1987b). *Science in action: How to follow engineers and scientists through society*. Milton Keynes, Open University Press.

Latour, B. (1988). *The pasteurization of France*. Cambridge, MA: Harvard University Press.

Latour, B. (1992). Where are the missing masses? Sociology of a few mundane artifacts. In W. Bijker, & J. Law, *Shaping technology/building society: Studies in sociological change*. Cambridge, MA: MIT Press.

Latour, B. (1996). *Aramis or the love of technology*. Cambridge, MA: Harvard University Press.

Law, J. (1992). Notes on the theory of the actor-network: Ordering, strategy and heterogeneity. *Systems Practice*, *5*(4): 379–393.

McMaster, T., Vidgen, R.T., & Wastell, D.G. (1997). *Towards an understanding of technology in transition. Two conflicting theories*. Proceedings of the Information Systems Research in Scandinavia, IRIS20 Conference, Hanko, Norway.

Mol, A., & Law, J. (1994). Regions, networks and fluids: Anaemia and social topology. *Social Studies of Science, 24***,** 641–671.

NHIMAC. (1999). Health on-line: A health information action plan for Australia. Canberra: NHIMAC.

Richards, B. (1999). Information technology in general practice. Canberra: General Practice Branch of the Commonwealth Department of Health and Aged Care.

Rogers, E.M. (1995). *Diffusion of innovations*. New York: The Free Press.

Tatnall, A. (2002). Modelling technological change in small business: Two approaches to theorising innovation. In S. Burgess, *Managing information technology in small business: Challenges and solutions* (pp. 83–97). Hershey, PA: Idea Group Publishing.

Tatnall, A., Everitt, P., Wenn, A., Burgess, S., Sellitto, C., & Darbyshire, P. (2004). A study of the adoption of information and communications technologies by rural general practitioners in Australia. In M.G. Hunter, & K.K. Dhanda, *Information systems: Exploring applications in business and government* (pp. 232–253). Washington, DC, The Information Institute.

Tatnall, A., & Gilding, A. (1999). *Actor-network theory and information systems research*. Proceedings of the 10th Australasian Conference on Information Systems (ACIS), Wellington, Victoria.

Tatnall, A., & Lepa, J. (2003). The Internet, e-commerce and older people: An actor-network approach to researching reasons for adoption and use. *Logistics Information Management, 16*(1), 56–63.

Wenn, A., Tatnall, A., Sellitto, C., Darbyshire, P., & Burgess, S. (2002). *A socio-echnical investigation of factors affecting IT adoption by rural GPs*. Proceedings of the IT in Regional Areas (ITiRA-2002) Conference, Rockhampton, Australia.

KEY TERMS

Actor-Network Theory: An approach to research in which networks associations and interactions among actors (both human and nonhuman) and are the basis for investigation.

Australian Divisions of General Practice Ltd: The peak national body representing 121 local groupings of GP, known as Divisions of General Practice, across Australia.

Essentialism: The philosophical view that for some specific entity there is a finite list of characteristics that determine its nature.

General Practitioner (GP): A General Practitioner (known in some countries as a Family Physician) is a medical doctor who provides primary care, treats acute and chronic illnesses, and provides preventive care and health education.

Innovation Diffusion: A theory of innovation in which the main elements are characteristics of the innovation itself, the nature of the communication channels, the passage of time, and the social system through which the innovation diffuses.

Innovation Translation: A theory of innovation in which, instead of using an innovation in the form it is proposed, potential adopters *translate* into a form that suits their needs.

Practice Manager: The administrative (i.e., non-medical) manager of a medical general practice.

Advances and Trends in Tissue Engineering of Heart Valves

Y. S. Morsi
Swinburne University of Technology, Australia

C. S. Wong
Swinburne University of Technology, Australia

S. S. Patel
Swinburne University of Technology, Australia

INTRODUCTION

Improvements in health care and treatment of diseases have led to an increase in life expectancy in developed countries. However, this achievement has also inadvertently increased the prevalence of chronic illnesses such as cardiovascular disease, adding to the growing burden of health care cost globally. Unfortunately, this trend is expected to escalate in the foreseeable future. Cardiovascular disease remains one of the main problems in contemporary health care worldwide, accounting for approximately one third of the world's total death (Poole-Wilson, 2005). This article focuses on a subgroup of cardiovascular disease known as valvular heart disease whereby abnormalities or malfunctions of the heart valves are detected. It is estimated that 93,000 valvular surgeries were conducted in the United States in 2002 (American Stroke Association & American Heart Association, 2005) and valve replacement surgeries accounted for 75% of the surgery performed for valvular defects in Australia and, of that, 56% were for aortic valves (Davies & Senes, 2003).

BACKGROUND

Currently, there are two types of artificial heart valves used in valve replacement surgeries: mechanical and tissue valves. However, these prostheses are not without limitations. Mechanical valves are usually made from pyrolytic carbon attached to a PET-covered metal such as titanium frame. Although more durable than tissue valves, patients implanted with mechanical valves are subjected to long-term complications such as thromboembolism, leading to a life-time administration of anti-coagulant (Bloomfield, Wheatley, Prescott, & Miller, 1991; Oxenham et al., 2003). Alternatively, tissue valves created from biological tissues from human or animal (porcine or bovine) may be used. While tissue valves do not require long-term anticoagulants, they undergo progressive deterioration such as calcification and tearing of cusps, leading to structural failure (Hammermeister, Sethi, Henderson, Oprian, Kim, & Rahimtoola, 1993; Schoen & Levy, 2005). Moreover, these clinically used prostheses are incapable of growth or remodelling. Hence, extensive research and development is being conducted worldwide to explore the potential of an emerging field, Tissue Engineering (TE), as a solution for addressing the shortcomings of current prosthesis used in valve replacement surgeries.

TE is a multidisciplinary area that amalgamates the principles of engineering and biological sciences to create functional tissue which can be ultimately used to repair, regenerate, or replace diseased or damaged parts of the body. A general approach of TE involves utilising temporary porous three-dimensional (3D) scaffolds to: (i) define the complex anatomical shape of the tissue, (ii) guide the proliferation and differentiation of seeded cells, and (iii) provide mechanical support for the cells (Ang, Leong, & Chua, 2006; Yang, Leong, Du, & Chua, 2001). While conventional techniques such as solvent casting and particulate leaching can be used to manufacture scaffolds, their applications to create scaffolds for heart valves are largely restricted by their limitations such as lack of mechanical strength and use of toxic organic solvents (Ang et al., 2006; Sachlos &

Czernuszka, 2003). The creation of scaffolds for heart valves require specific characteristics and precision of the tissue to be captured and ideally these scaffolds should be tailored to individual patients.

RAPID PROTOTYPING TECHNOLOGIES

Recently, an increased interest has been generated for Rapid Prototyping (RP) techniques as powerful tools for fabrication of scaffolds. RP techniques may be able to address some of the limitations encountered in the conventional techniques. There are three types of RP techniques discussed in this article: fused deposition modeling (FDM), 3D printing, and bioprinting.

FDM is a material deposition process which uses a computer-aided design (CAD) model to generate 3D scaffolds (Masood, Singh, & Morsi, 2005). The scaffolds are generated through the extrusion of thin rods of molten polymer using a computer-controlled XYZ robotic dispenser (Figure 1). The layers of polymer are deposited in an interconnected manner, thus improving the mechanical stability of the scaffold. FDM enables complex yet accurate characteristics to be reconstructed from CT scans, which leads to the ability to create scaffolds customised to patients' needs. Scaffolds demonstrating the complex geometry of the aortic valve which incorporated the exact dimensions of the sinuses of Valsalva (required to preserve the flow characteristics of the valve) were successfully manufactured using FDM (Figure 2) (Morsi & Birchall, 2005). This technique offers a high degree of control over the shape, pore interconnectivity, and porosity of scaffolds as individual process parameters can be defined and improved (Ang et al., 2006). A high resolution of 250 μm can be achieved with the FDM. An added advantage of the FDM technique is that the process does not utilise toxic solvents and porogens for the manufacturing of scaffolds (Leong, Cheah, & Chua, 2003; Yang, Leong, Du, & Chua, 2002). The flexibility of this technique lends itself to produce scaffolds of varying designs and complexity, thus expanding its application to other areas of TE aside from heart valves.

The 3DP is a layered fabrication process whereby sliced 2D profile of a model determined by CAD file is transform to a STL file and printed on a fresh layer of powder via deposition of a suitable binder from the inkjet printhead (Figure 3). The powders used are natural polymers such as starch, dextran, and gelatin in conjunction with water-based binder so that the scaffolds can be used in medical applications, as toxic solvents can be omitted from the manufacturing process. The 2D profiles are then successively printed on freshly laid layers of powder until the entire scaffold is printed. These printed layers are held together through

Figure 1. Diagram demonstrating the fabrication process of FDM. The ABS filament is fed through the FDM head via the drive wheels and melted at the appropriate temperature in the liquefier. The polymeric fibres of 250 μm are then deposited layer by layer in an XYZ direction to produce a 3D scaffold.

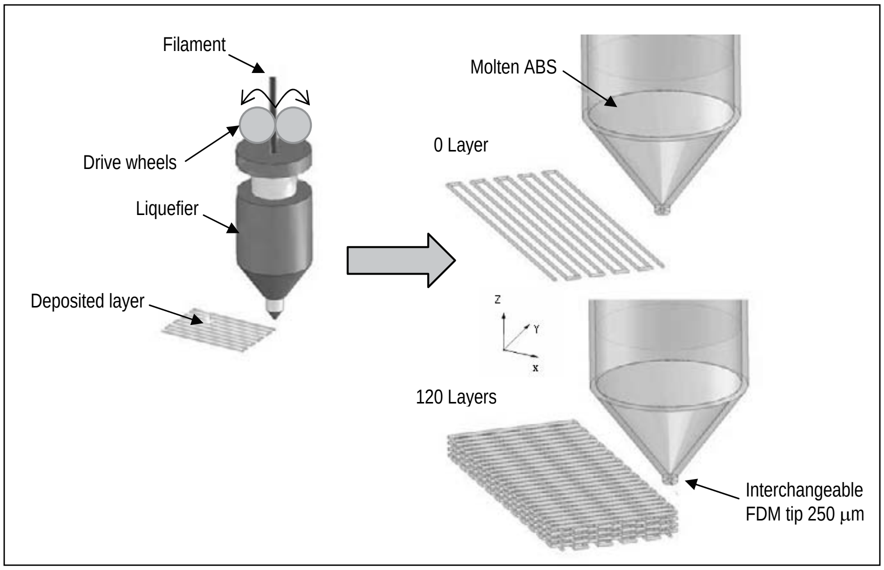
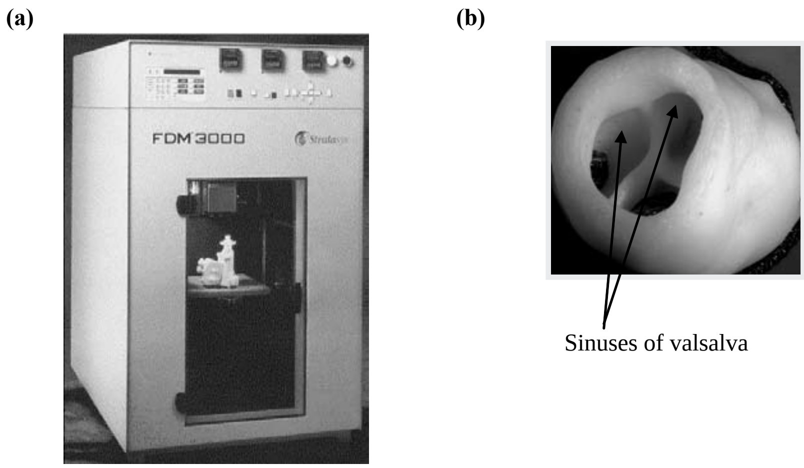


Figure 2. (a) FDM machine which produces 3D scaffolds using a deposition process; (b) model of an ovine aortic heart valve including the sinuses of Valsalva manufactured using FDM



the binder and the completed model then undergoes post-processing to improve its water resistance and mechanical strength. This is because models created via 3DP are weak and easily damaged. The post-processing involves the printed scaffolds being dried at 100°C and impregnated with molten wax or polymeric solutions such as polycaprolactone and poly-L-lactide acid, followed by drying at room temperature (Lam, Mo, Teoh, & Hutmacher, 2002). Shrinkage of scaffolds can occur during post-processing and the proportion is dependent on the shape of the scaffold. Scaffolds with circular pore designs have been shown to have the highest shrinkage (Lam et al., 2002). The 3DP techniques enable scaffolds to be made at a fast pace and are cost-effective as compared to the FDM process. This technique is advantageous for producing scaffolds that do not require precision with regards to dimensions and mechanical strength, for example, as a prototype for conceptualisation purposes.

Bioprinting is a form of RP technology that incorporates biological components such as living cells into the scaffold produced. Typically, standard inkjet printers are modified so that cells suspended in liquid can be sprayed onto biodegradable, thermosensitive gel based on computer generated templates (Mironov, Boland, Trusk, Forgacs, & Markwald, 2003; Ringeisen, Othon, Barron, Young, & Spargo, 2006). The gel is then printed layer by layer to create the 3D model. Currently, a resolution of approximately 100 μm can be achieved (Mironov, Reis, & Derby, 2006). Recent improvements on the bioprinting technology have enabled aggregate of cells to be deposited in droplets via a micropipette within the printing head, as shown in Figure 4 (Jakab, Damon, Neagu, Kachurin, & Forgacs, 2006). This would increase the density of cells being "printed" onto the scaffold and possibly accelerating the formation of tissue.

It is essential that the printing process does not affect the viability of the cells; therefore, generation of scaffold within a short period of time is advantageous. Other factors that need to be considered include the heat generated by the inkjet system, shear stress experienced by cells during ejection from the printer head, and the impact upon contact with the scaffold. Furthermore, the printing process should not be toxic to the cells and the scaffold generated should be well interconnected. Printed chicken heart cells have been shown to survive the process and after 19 hours demonstrated normal heart cell characteristics such as beating synchronously (Jakab et al., 2006). An electrostatically driven printing system that generated less heat than the conventional system was recently demonstrated to be conducive to bovine endothelial cells (Nakamura et al., 2005). Bioprinting has the potential to create viable tissue such as blood vessels by, for example, printing consecutive circular layers of smooth muscle cells lined with endothelial cells to form a tube serving as replacements for arteries and veins. Since the cells are already incorporated in or onto the scaffold, the time required for investigating cellular growth and characteristics such as functionality *in vitro* can be reduced. A better understanding of the mechanism will ensure effective implementation of the technology to produce 3D scaffolds.

Figure 3. A schematic diagram of the 3D printing system using an inkjet-based printing technique to produce a 3D model of trileaflet heart valve

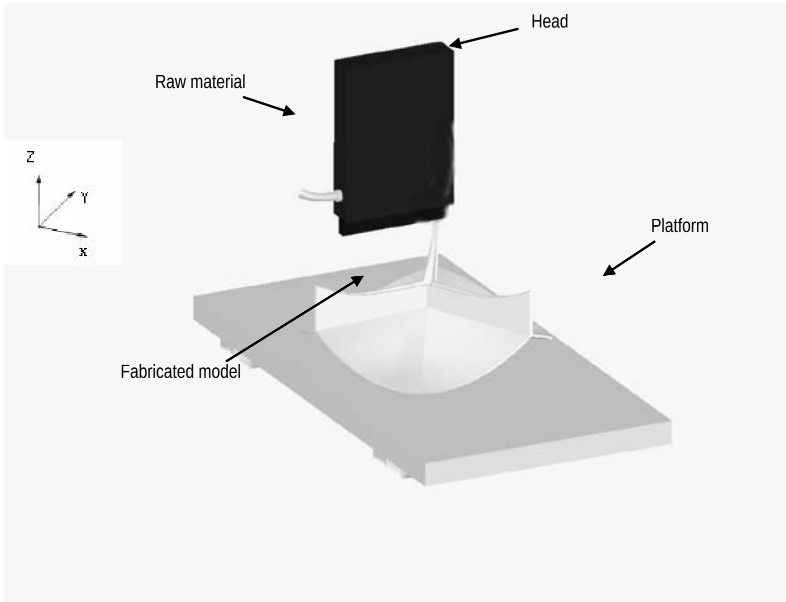
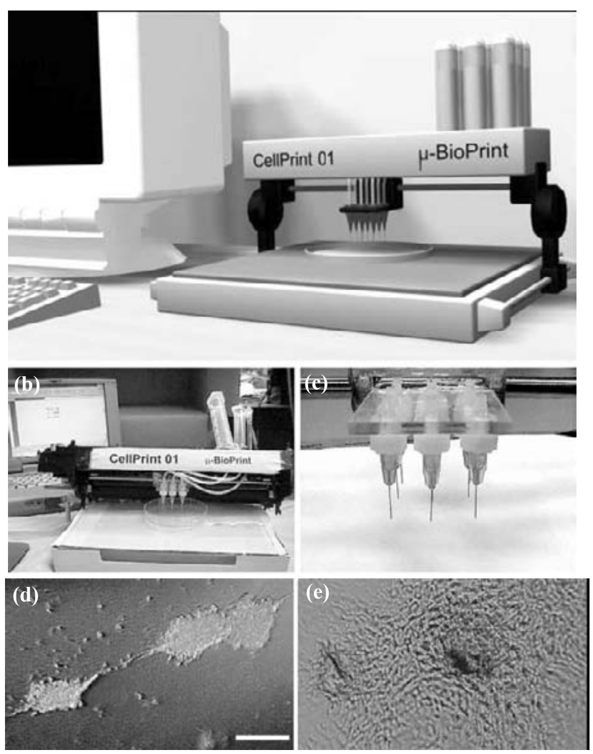


Figure 4. The bioprinting system. (a) A schematic diagram showing the creation of 3D tube via bioprinting whereby droplets of cells are deposited layer by layer onto a solidifying gel. (b) Cell printer showing a part of the print head with 9 nozzles. (c) The print head is attached to a bidirectional parallel cable which is connected to the printer, controlled by a microchip. (d) Endothelial cells printed on collagen before fusion. e) Aggregates of endothelial cells after fusion. Reproduced with permission.[1]



RP techniques are increasingly being used in the area of prosthetics. While TE of organs requires the creation of an anatomically correct scaffold, the mechanical properties of the scaffold also play an important role in tissue regeneration. The mechanical strength of the scaffold should be comparable to that of the organ or tissue of interest. For example, scaffolds for heart valves must be able to withstand the haemodynamic pressures experienced *in vivo*. Factors that influences mechanical properties include porosity of the scaffolds, strength and elasticity of the materials used (Yang et al., 2001). The technologies of RP have the flexibility to produce scaffolds with varying porosity to accommodate specific requirements. This is because the parameters of the RP techniques can be modified and controlled using softwares that drive the machine, giving rise to numerous possibilities. The flexibility offered by RP technologies is important as the optimal porosity differs between cell types with respect to cell attachment and proliferation, which in turn affect the mechanical properties required. Consequently, a balance or compromise is needed between these factors to produce a scaffold that can carry out a particular function satisfactorily.

CURRENT RESEARCH AND FUTURE TRENDS

RP is a technology that is developing rapidly, diverging from its traditional role in engineering to being used as a powerful tool in medicine, for example, in the area of prosthetics, surgical planning, and medical instrumentation. The transition from a technology mainly used in engineering to medicine was eased by the fact that computed tomography (CT) techniques used to scan various parts of the body operate via similar layer-based technologies as RP. Data obtained from the CT scans can be interpreted and translated to the RP system accurately.

RP techniques have been used in hip replacements and maxillofacial prostheses where models are created so that the resulting implant can be fitted onto the patient precisely (Eggbeer, Bibb, & Evans, 2006; Sanghera, Naique, Papaharilaou, & Amis, 2001; Sykes, Parrott, Owen, & Snaddon, 2004). Complex and challenging tasks such as sculpting an ear cast have also turn to RP so that an accurate model of the patient's ear can be generated in a time-efficient manner (Al Mardini, Ercoli, & Graser, 2005). Moreover, these techniques have been implemented to aid surgeons in their surgical planning for reconstruction procedures and act as guides during the procedure so that optimal outcomes, both clinically and aesthetically, were obtained, as well as for education purposes such as improving the understanding of trainees (Muller, Krishnan, Uhl, & Mast, 2003; Poukens, Haex, & Riediger, 2003; Toso et al., 2005).

Research into TE of heart valves has utilised RP in an attempt to address the aforementioned limitations of currently used prosthetic heart valves. Generally, a polymeric scaffold of heart valves is produced using RP and then seeded with cells of interest such as endothelial cells, valvular cells, and stem cells (Morsi, Birchall, & Rosenfeldt, 2004; Sodian et al., 2005). Research into stem cells in the recent years has presented itself as an exciting area in TE due to the potential of being undifferentiated cells that have the capability of developing into cells with specialised functions such as heart, liver and kidney cells. This suggests that given the right conditions, stem cells seeded onto scaffolds of heart valves may develop into valvular cells required to produce a functioning organ. However, the mechanisms of stem cells are yet to be fully understood and it is beyond the scope of this article to explore this area further.

Currently, RP technology in the form of FDM does not support the direct fabrication of scaffolds using biocompatible and biodegradable polymers. This is because initially, FDM machines utilise acrylonitrile butadiene styrene exclusively as its source of polymer. Subsequent modifications to the system enabled limited polymers, for example, polycaprolactone to be used by the FDM machine to create scaffolds for TE of bone (Chim et al., 2006). The success of modifying FDM systems to accommodate polymers other than acrylonitrile butadiene styrene has encouraged research into these systems to support other biocompatible and/or biodegradable polymers so that the models produced are clinically applicable. However, through FDM technology, a precise mold of a heart valve can be fabricated which in turn facilitates the creation of scaffold with polymer of one's choice. This indirect technique enables the design of heart valve scaffolds to be optimised, geometrically and biologically. The complex geometry of heart valves was successfully reproduced by RP technology and scaffolds of trileaflet heart valve created using synthetic polymers such as poly-4-hydroxybutyrate were demonstrated to be capable of opening and closing synchronously (Sodian et al., 2002). Other polymers such as polyglycolic and polylactic acid have also been used to create scaffolds for heart valves (Kim et al., 2000; Shinoka, 2002). Moreover, it is feasible to create heart valve scaffolds from natural polymers such as collagen using indirect RP techniques (Rothenburger et al., 2002; Sachlos, Reis, Ainsley, Derby, & Czernuszka, 2003; Taylor, Sachlos, Dreger, Chester, Czernuszka, & Yacoub, 2006).

CONCLUSION

RP is being increasingly used in medical research due to its flexibility in creating a variety of complex shapes which can be custom made to the patient's specifications. This ensures that implants will be optimally positioned in individual patients, thus improving the quality of treatment. Manufacturing scaffolds using RP technology also enable the transformation from design to a 3D model in a time-efficient and cost effective manner. Additionally, the operation of RP systems requires minimal human resources as most of the processes are automated. The ultimate goal is to refine existing RP technology so that it is applicable to medical research whereby scaffolds of heart valves fabricated from biocompatible and biodegradable polymer can be achieved using direct RP techniques. Such an achievement will not only produce heart valve scaffolds that are anatomically correct but ones that will degrade as extracellular matrix and tissue gradually takes over the scaffold, taking it a step closer to the generation of a true living tissue.

ACKNOWLEDGMENT

The authors would like to thank Mr. Toby Lai for his assistance in creating the diagrams.

REFERENCES

Al Mardini, M., Ercoli, C., & Graser, GN. (2005). A technique to produce a mirror-image wax pattern of an ear using rapid prototyping technology. *J. Prosthet. Dent, 94*(2), 195-198.

American Stroke Association & American Heart Association. (2005). Heart disease and stroke statistics - 2005 update (November, 21, 2006). Retrieved January 30, 2008, from http://www.americanheart.org/downloadable/heart/1105390918119HDSStats2005Update.pdf

Ang, K.C., Leong, K.F., & Chua, CK. (2006). Investigation of the mechanical properties and porosity relationships in fused deposition modelling-fabricated porous structures. *Rapid Prototyping Journal, 12*(2), 100-105.

Bloomfield, P., Wheatley, D.J., Prescott, R.J., & Miller, HC. (1991). Twelve-year comparison of a Bjork-Shiley mechanical heart valve with porcine bioprostheses. *N Engl J Med, 324*(9), 573-579.

Boland, T., Mironov, V. Gutowska, A., Roth, E.A., & Markwald, R.R. (2003). Cell and organ printing 2: Fusion of cell aggregates in three-dimensional gels. *Anat Rec A Discov Mol Cell Evol Biol, 272*(2), 497-502.

Chim, H., Hutmacher, D.W., Chou, A.M., Oliveira, A.L., Reis, R.L., Lim, T.C., et al. (2006). A comparative analysis of scaffold material modifications for load-bearing applications in bone tissue engineering. *Int J Oral Maxillofac Surg, 35*(10), 928-934.

Davies, J., & Senes, S. (2003). Cardiac surgery in Australia 1999. Retrieved January 30, 2008, from http://www.aihw.gov.au/publications/index.cfm/title/8582

Eggbeer, D., Bibb, R., & Evans, P. (2006). Toward identifying specification requirements for digital bone-anchored prosthesis design incorporating substructure fabrication: A pilot study. *Int J Prosthodont, 19*(3), 258-263.

Hammermeister, K.E., Sethi, G.K., Henderson, W.G., Oprian, C., Kim, T., & Rahimtoola, S. (1993). A comparison of outcomes in men 11 years after heart-valve replacement with a mechanical valve or bioprosthesis. Veterans Affairs Cooperative Study on Valvular Heart Disease. *N Engl J Med, 328*(18), 1289-1296.

Jakab, K., Damon, B., Neagu, A., Kachurin, A., & Forgacs, G. (2006). Three-dimensional tissue constructs built by bioprinting. *Biorheology, 43*(3-4), 509-513.

Kim, W.G., Park, J.K., Park, Y.N., Hwang, C.M., Jo, Y.H., Min, B.G., et al. (2000). Tissue-engineered heart valve leaflets: An effective method for seeding autologous cells on scaffolds. *Int J Artif Organs, 23*(9), 624-628.

Lam, C.X.F., Mo, X.M., Teoh, S.H., & Hutmacher, DW. (2002). Scaffold development using 3D printing with a starch-based polymer. *Materials Science and Engineering, C*(20), 49-56.

Leong, K.F., Cheah, C.M., & Chua, CK. (2003). Solid freeform fabrication of three-dimensional scaffolds for engineering replacement tissues and organs. *Biomaterials*, *24*(13), 2363-2378.

Masood, S.H., Singh, J.P., & Morsi, Y. (2005). The design and manufacturing of porous scaffolds for tissue engineering using rapid prototyping. *The Int J Adv Manufac Tech, 27*(3-4), 415-420.

Mironov, V., Boland, T., Trusk, T., Forgacs, G., & Markwald, RR. (2003). Organ printing: Computer-aided jet-based 3D tissue engineering. *Trends Biotechnol, 21*(4), 157-161.

Mironov, V., Reis, N., & Derby, B. (2006). Review: Bioprinting: A beginning. *Tissue Eng, 12*(4), 631-634.

Morsi, Y.S., & Birchall, IE. (2005). Tissue engineering a functional aortic heart valve: An appraisal. *Future Cardiol, 1*(3), 405-411.

Morsi, Y.S., Birchall, I.E., & Rosenfeldt, FL. (2004). Artificial aortic valves: An overview. *Int J Artif Organs, 27*(6), 445-451.

Muller, A., Krishnan, K.G., Uhl, E., & Mast, G. (2003). The application of rapid prototyping techniques in cranial reconstruction and preoperative planning in neurosurgery. *J Craniofac Surg, 14*(6), 899-914.

Nakamura, M., Kobayashi, A., Takagi, F., Watanabe, A., Hiruma, Y., Ohuchi, K., et al. (2005). Biocompatible inkjet printing technique for designed seeding of individual living cells. *Tissue Eng, 11*(11-12), 1658-1666.

Oxenham, H., Bloomfield, P., Wheatley, D.J., Lee, R.J., Cunningham, J., Prescott, R. J., et al. (2003). Twenty year comparison of a Bjork-Shiley mechanical heart valve with porcine bioprostheses. *Heart, 89*(7), 715-721.

Poole-Wilson, P. (2005). The prevention of cardiovascular disease worldwide: Whose task and WHO's task? *Clin Med, 5*(4), 379-384.

Poukens, J., Haex, J., & Riediger, D. (2003). The use of rapid prototyping in the preoperative planning of distraction osteogenesis of the cranio-maxillofacial skeleton. *Comput Aided Surg, 8*(3), 146-154.

Ringeisen, B.R., Othon, C.M., Barron, J.A., Young, D., & Spargo, BJ. (2006). Jet-based methods to print living cells. *Biotechnol J, 1*(9), 930-948.

Rothenburger, M., Volker, W., Vischer, J.P., Berendes, E., Glasmacher, B., Scheld, H.H., et al. (2002). Tissue engineering of heart valves: Formation of a three-dimensional tissue using porcine heart valve cells. *Asaio J, 48*(6), 586-591.

Sachlos, E., & Czernuszka, JT. (2003). Making tissue engineering scaffolds work. Review: The application of solid freeform fabrication technology to the production of tissue engineering scaffolds. *Eur Cell Mater, 5*, 29-39; Discussion 39-40.

Sachlos, E., Reis, N., Ainsley, C., Derby, B., & Czernuszka, JT. (2003). Novel collagen scaffolds with predefined internal morphology made by solid freeform fabrication. *Biomaterials, 24*(8), 1487-1497.

Sanghera, B., Naique, S., Papaharilaou, Y., & Amis, A. (2001). Preliminary study of rapid prototype medical models. *Rapid Prototyping Journal, 7*(5), 275-284.

Schoen, F.J., & Levy, RJ. (2005). Calcification of tissue heart valve substitutes: Progress toward understanding and prevention. *Ann Thorac Surg, 79*(3), 1072-1080.

Shinoka, T. (2002). Tissue engineered heart valves: Autologous cell seeding on biodegradable polymer scaffold. *Artif Organs, 26*(5), 402-406.

Sodian, R., Fu, P., Lueders, C., Szymanski, D., Fritsche, C., Gutberlet, M., et al. (2005). Tissue engineering of vascular conduits: Fabrication of custom-made scaffolds using rapid prototyping techniques. *Thorac Cardiovasc Surg, 53*(3), 144-149.

Sodian, R., Loebe, M., Hein, A., Martin, D.P., Hoerstrup, S.P., Potapov, E.V., et al. (2002). Application of stereolithography for scaffold fabrication for tissue engineered heart valves. *Asaio J, 48*(1), 12-16.

Sykes, L.M., Parrott, A.M., Owen, C.P., & Snaddon, DR. (2004). Applications of rapid prototyping technology in maxillofacial prosthetics. *Int J Prosthodont, 17*(4), 454-459.

Taylor, P.M., Sachlos, E., Dreger, S.A., Chester, A.H., Czernuszka, J.T., & Yacoub, MH. (2006). Interaction of human valve interstitial cells with collagen matrices manufactured using rapid prototyping. *Biomaterials, 27*(13), 2733-2737.

Thompson, R.P., Reckova, M., de Almeida, A., Bigelow, M.R., Stanley, C.P., & Spruill, J.B. et al. (2003). *The oldest, toughest cells in the heart.* Novartis Found Symp. (pp. 157-174).

Toso, F., Zuiani, C., Vergendo, M., Salvo, I., Robiony, M., Politi, M., et al. (2005). Usefulness of computed tomography in pre-surgical evaluation of maxillofacial pathology with rapid prototyping and surgical pre-planning by virtual reality. *Radiol Med (Torino), 110*(5-6), 665-675.

Yang, S., Leong, K.F., Du, Z., & Chua, CK. (2001). The design of scaffolds for use in tissue engineering. Part I. Traditional factors. *Tissue Eng, 7*(6), 679-689.

Yang, S., Leong, K.F., Du, Z., & Chua, CK. (2002). The design of scaffolds for use in tissue engineering. Part II. Rapid prototyping techniques. *Tissue Eng, 8*(1), 1-11.

KEY TERMS

Biocompatible Polymer: Polymer that does not produce toxin or harmful products and stimulate an immune response in biological systems. This is essential so that during implantation, the material does not induce a rejection response.

Biodegradable Polymer: Polymer that can be broken down into nontoxic by-products through biological processes. Biodegradation can occur physically or chemically via enzymes or bacteria. Examples of biodegradable polymers are polyglycolic acid and polycaprolactone.

Cell Proliferation: An increase in the number of cells due to cell growth and cell division. In the area of tissue engineering, proliferation of cells would enable cells attached to the scaffold to gradually cover the surface area of interest, producing a layer of live cells.

Computer Aided Design: Also known as CAD. A computer based software program which is commonly used to assist engineers and architects in designing detailed 3D models. CAD is also a useful tool for conceptualisation of designs and visualisation of models prior to production.

Fused Deposition Modeling: Commonly known as FDM. A type of RP technique. FDM produces 3D scaffolds by heating the material in the extruder and depositing them in a horizontal and vertical direction (XYZ direction), which is controlled by CAD.

Rapid Prototyping Technique: An advanced manufacturing technique that produces 3D models in an automated manner through sequential layering of the material. Models can be created in a time frame ranging from 3 to 72 hours, depending on size.

Scaffold: A supporting structure for cells and tissue to grow on; a scaffold will provide the physical structure, which guides the cells to grow to the correct anatomical shape.

ENDNOTE

1 **(Figures 4a-e):** Reprinted from *Trends in Biotechnology, 21*(4); Mironov, V. et al., *Organ printing: Computer-aided jet-based 3D tissue engineering*, 157-161, 2003, with permission from Elsevier. **(Figure 4a):** Reproduced with permission from Thompson, R.P. et al. (2003). *The oldest, toughest cells in the heart.* Copyright John Wiley & Sons Limited. **(Figure 4d & e):** Reprinted with permission of Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Inc. Cell and organ printing 2: Fusion of cell aggregates in three-dimensional gels, Boland T. et al., (2003). *The anatomical record—Part A.*

Advances in Bone Tissue Engineering to Increase the Feasibility of Engineered Implants

Neelima Vidula
Northwestern University, USA

Jessy J. Mouannes
Northwestern University, USA

Nadia Halim
University of Illinois at Chicago, USA

Shadi F. Othman
University of Illinois at Chicago, USA

INTRODUCTION

Millions of patients experience bone loss as a result of degenerative disease, trauma, or surgery (Xu, Othman, Hong, Peptan, & Magin, 2005). Healthy bone tissue constantly regenerates itself and remodels its architecture to meet the mechanical demands imposed on it, as described by Wolff's "Law of Bone Remodeling" (Wolff, 1986). However, this capacity is severely limited when there is insufficient blood supply, mechanical instability, or competition with highly proliferating tissues (Pinheiro & Gerbei, 2006). Furthermore, severe bone losses can be detrimental to individuals, because they reduce the bone's ability to remodel, repair, and regenerate itself (Luo et al., 2005; Nordin & Franklin, 2001), ultimately resulting in the deterioration of a patient's health, and, in some instances, death (Luo et al., 2005).

Because the repercussions of bone loss are severe, it is important to replace lost bone in patients. The current *gold standard* for specific-site structural and functional bone defect repair is autologous bone grafts (Mauney, Volloch, & Kaplan, 2005) or autografts. While autografts do not present the problem of immune rejection, since the bone tissue is being transplanted from another region of the patient's own body (Rahaman & Mao, 2005), they present certain complications such as significant donor site morbidity (death of tissue remaining in the region from which the donor tissue was removed), infection, malformation, and subsequent loss of graft function (Mauney et al., 2005).

Another established and widely employed technique for the treatment of bone loss is the transplantation of allograft bone or bone tissue from a donor (Mauney et al., 2005). Although allograft bone is effective in treating bone loss, there are several common problems associated with it: first, a compatible donor must be found (Jones, Erhenfried, & Hench, 2006) in order to minimize the possibility of immune rejection by the patient; second, there is a risk of potential disease transmission from the donor to the patient; third, donor site morbidity can occur (Jones et al., 2006); and, finally, there is a limited supply of donor tissue (Mauney et al., 2005). Therefore, patients often experience long waiting periods before receiving the transplant, due to the scarcity of tissue donors, and this can exacerbate bone tissue loss (Jones et al., 2006).

The development of the field of bone tissue engineering has expanded the solutions available to the problem of bone tissue loss. Arguably, implants developed via tissue engineering applications may be a more viable solution to the problem of bone loss than conventional solutions. In contrast with transplants, such implants are not subject to patient-donor tissue biocompatibility issues, because donor tissue is unnecessary. Also, morbidity of the site of extracted tissue is not as great of a problem, since implants can be developed from less tissue. Additionally, implants are generally more readily available to patients than transplants, which reduces the time for initiation of bone loss treatment (Jones et al., 2006). Therefore, bone tissue engineering may very well be the future gold standard treatment for bone loss.

BACKGROUND

One basic scheme of the bone tissue engineering process currently employed is illustrated in Figure 1. Briefly, mesenchymal stem cells are obtained from the patient, generally from the bone marrow (Stock & Vacanti, 2001). After a period of cellular expansion, the cells are seeded on biodegradable and biocompatible scaffolds (Stock & Vacanti, 2001). Poly-DL-lactic-co-glycolic acid (PLGA), gelatin, and collagen scaffolds are frequently employed as surfaces for bone tissue development (Wu, Shaw, Lin, Lee, & Yang, 2006; Xu et al., 2005; Zhang et al., 2006). These scaffolds are supplemented with bone differentiation promoting factors such as bone-morphogenic protein, dexamethasone, and ascorbate-2-phosphate that enable the stem cells to differentiate into osteoblasts (bone-forming cells) (Kim et al., 2005). After a substantial period of culturing, implantation of the scaffold into the patient occurs, leading to bone restoration (Xu et al., 2005).

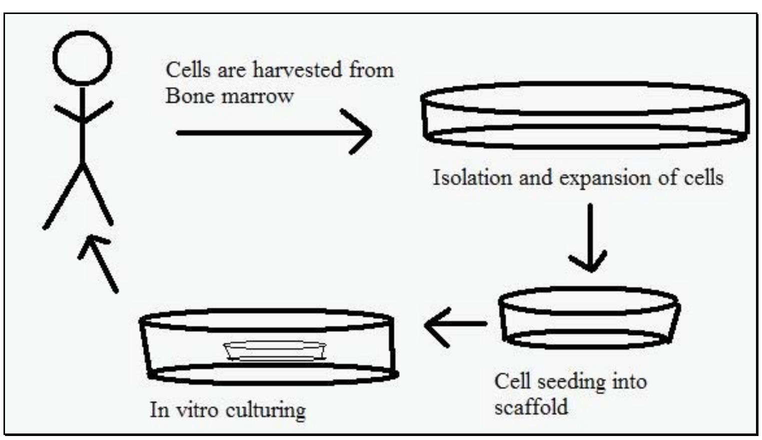
Although this process has the potential to treat bone loss, it is far from optimal. Formation of engineered bone tissue currently takes several weeks (at least 3 to 4 weeks), resulting in extensive waiting periods for patients (Cartmell et al., 2005). Since time is of the essence for patients with bone loss, reducing the culture time of stem cells is necessary for implants to be effective. In addition, a portion of the engineered tissue is destroyed during invasive histological assessment conducted to confirm the formation of bone tissue. This form of assessment can further increase patient waiting periods, as the portion of engineered tissue used for testing is no longer available for implantation. A need exists for a bone tissue engineering process that overcomes these problems.

Reducing the culture time of stem cells is necessary to increase the effectiveness of engineered implants. The use of electrical and mechanical stimulation devices to accelerate stem cell differentiation into osteoblasts has been implicated.

Electric and electromagnetic fields may be effective in accelerating stem cell differentiation. Aaron, Ciombor, and Simon (2004) demonstrated that electric and electromagnetic fields can accelerate bone formation and healing, particularly in osteotomies and spine fusions, both *in vivo* and *in vitro*; osteotomy is a procedure in which bone is surgically cut to improve alignment (ASBA Glossary, 2005). Electric fields can be generated either invasively in bone, by placing electrodes directly into the healing site, or noninvasively, through capacitive or inductive coupling. Osteogenesis, or bone formation, is usually stimulated at the cathode (Aaron et al., 2004).

"Electrical Properties of Bone" by Lakes (2005) describes another type of stimulation that can be achieved by means of a piezoelectric actuator. It has been demonstrated by several researchers that bone is piezoelectric. The piezoelectric nature of bone indicates that any mechanical stress applied to bone can produce

Figure 1. A current process of bone tissue engineering is depicted



an electric polarization of the tissue, and any electric field applied to bone can cause mechanical strain of the tissue. Piezoelectric effects occur in the kilo-hertz range, well above that of physiologically significant frequencies.

Pilla (2002) describes how low-intensity electromagnetic and mechanical modulations of bone growth and repair are equivalent. To explain that, there is a time-varying electric field, E(t), associated with both types of stimuli, which serves as a main messenger regulating cellular activity, therefore acting as a growth stimulus. This electric field can be either directly induced with electric and electromagnetic devices, or indirectly induced by an applied mechanical stress. One way to generate the latter is to use a piezoelectric actuator as described previously, and another way, which has also been demonstrated to work effectively for both *in vivo* and *in vitro* bone formation and repair by Nolte et al. (2001), is to use low-intensity ultrasound waves. Bone repair is significantly enhanced by both electromagnetic fields and ultrasound signals (Pilla, 2002).

In tissue engineering involving osteoblasts seeded onto scaffolds and grown *in vitro*, a recent study by Cartmell et al. (2005) has proven that mechanical forces, when applied within a certain dosage, influence osteoblast activity, by accelerating their growth. The solution for stimulation used in this study was "magnetic particles" that were coated with RGD peptides (adsorption-promoting proteins), and then attached to the human osteoblasts. Exposure to a magnetic field would cause the magnetic particles to move or vibrate, thus creating mechanical stimulation.

The various stimulation methods cited above demonstrate the recent trend towards developing technologies that improve established tissue engineering processes. Yet, the optimized usage of such stimulation devices has yet to occur (Zhang et al., 2006), and is dependent upon further research.

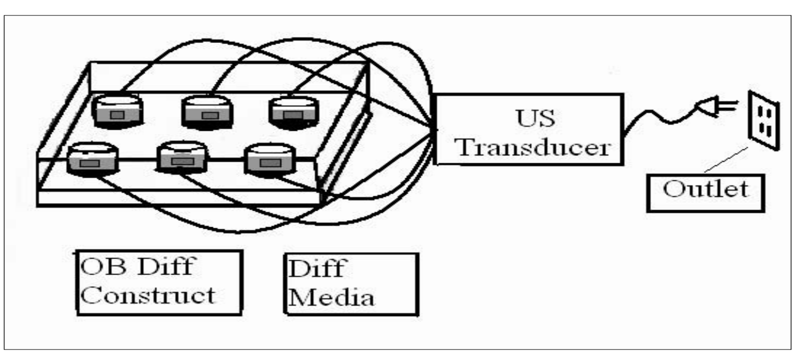
DEVELOPMENT OF A FEASIBLE BONE TISSUE ENGINEERING PROCESS

The development of a feasible bone tissue engineering process calls for a combination of a noninvasive stimulation device to reduce stem cell culture time, and a noninvasive method for monitoring the growth of the engineered constructs.

Of the various methods available to reduce the culture time of engineered bone tissue, ultrasound (US) stimulation may be the most promising one. This is because US is well known to be a noninvasive technique (Buckwalter & Brown, 2004), which is very relevant to bone tissue engineering because it insures the integrity of the engineered constructs, which are generally small and delicate. Moreover, previous studies have indicated that low-intensity pulsed US, administered with a dose as short as 20 minutes per day, activated ossification *in vitro* via a direct effect on osteoblasts and ossifying cartilage, after other animal and clinical studies showed that low-intensity US accelerated bone healing *in vivo* (Xu et al., 2005).

Magnetic resonance microscopy (MRM) has been proven to be an efficient, quantitative technique that could be applied to noninvasively monitor the bone tissue engineering process. This technique entails the application of a radiofrequency (RF) pulse to excite a tissue or specimen at resonance, and the acquisition of signal in the form of RF energy during subsequent relaxation of tissue magnetization. This method provides quantitative parameters that are directly dependent on the tissue properties, such as the spin-spin T_2

Figure 2. Ultrasound stimulation setup



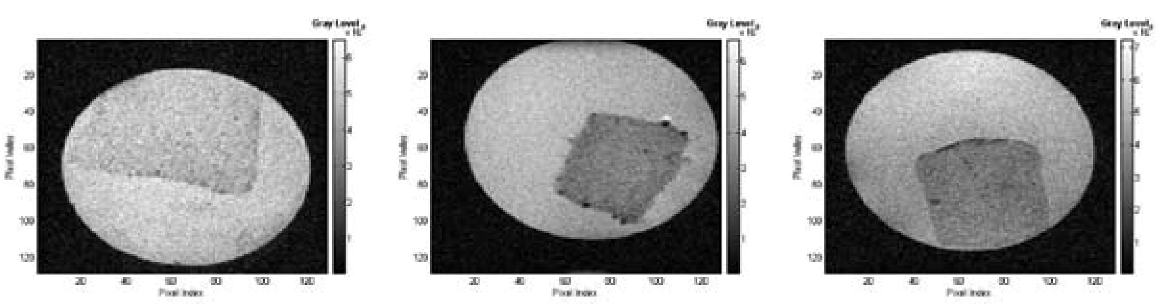
relaxation time, which describes the time it takes for one of the magnetization components, the transverse component, to relax back to its equilibrium condition following excitation (Nishimura, 1996). Currently, magnetic resonance imaging (MRI) is widely used *in vivo* to assess connective tissue degeneration (Cartmell et al., 2005), and also to investigate the regeneration of engineered tissue (Carnegie Mellon University, 2005). A recent study showed that MRM can be used to monitor osteogenesis in tissue-engineered constructs (Xu et al., 2005), indicating that MRM and related technologies may be feasible for replacing the conventional, invasive histological assessment of engineered bone tissue.

With these considerations, in a recent work (Moinnes, Vidula, Halim, & Othman, 2006), we explored the usage of US stimulation for reducing the culture time of stem cells, and also the use of MRM for monitoring the growth of engineered bone tissue. Working off the conventional bone tissue engineering scheme, we isolated mesenchymal stem cells from human bone marrow, and then seeded these cells on collagen scaffolds; these scaffolds were grown *in vitro* for 3 weeks and were divided into three groups: control constructs, differentiated, nonstimulated constructs, and differentiated, ultrasound stimulated constructs, with the latter two groups exposed in culture to bone differentiation promoting factors. Ultrasound (Figure 2) was administered using a sonic accelerated fracture-healing system device and transducers delivering pulsed ultrasound waves with an intensity of 30 mW/cm^2, operating frequency of 1.5MHz, pulse width of 200 μsec, and pulse repetition rate of 1 kHz for 20 minutes per day throughout the growth period.

At 1 week intervals, we conducted both MRM and histological assessment of growth of the engineered constructs. MRM experiments were conducted at 11.74 T (500 MHz for protons) using a 56 mm vertical bore magnet (Oxford Instruments, Oxford, UK), and a Bruker DRX Avance Spectrometer (Bruker Instruments, Billerica, MA, USA) controlled by a Silicon Graphics SGI2 workstation (Mountain View, CA, USA). MR images were acquired using a Bruker Micro 5 imaging probe with triple axis gradients (maximum strength 2 T/m), and a 10 mm diameter RF saddle coil was used to transmit/receive the nuclear magnetic resonance signals. Our MRM analysis included the determination of the spin-spin relaxation time (T_2) for specific regions of interests localized at the periphery of each construct. Histological assessment, for correlation with the MR T_2 determination, included hematoxylin and eosin staining for the detection of cell nuclei, von Kossa staining for examination of mineralization due to calcium deposition by newly formed bone, and osteocalcin (bone matrix protein) for confirmation of bone tissue differentiation.

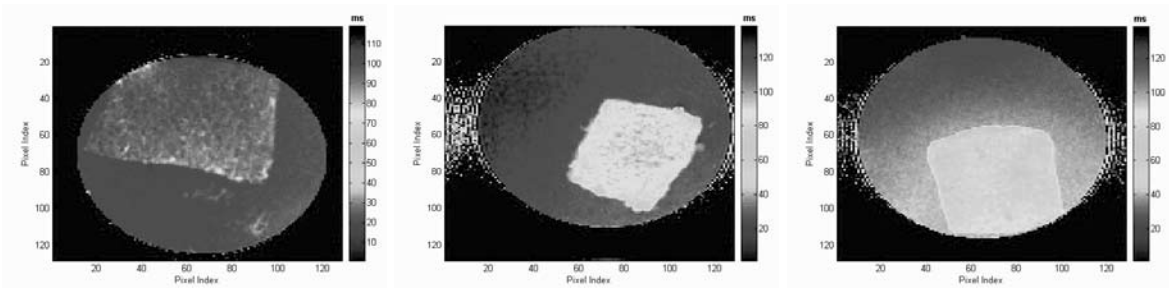
Using these methods, we drew several conclusions, which may help justify the usage of ultrasound stimulation and MRM construct monitoring in enhancing the current scheme for bone tissue engineering. First, we demonstrated a good correlation between the MR and histological/immunocytochemical results; we saw a reduction in the T_2 relaxation time values for differentiated ultrasound stimulated constructs after 2 weeks of growth, indicating an increase in construct stiffness due to bone formation, which we correlated with increased bone tissue formation determined through histological analysis (Figures 3, 4, 5, & 6). Thus, we demonstrated the feasibility of MRM for replacing histological assessment of the growth of engineered tissue. Moreover, we established the effectiveness of low intensity pulsed

Figure 3. Variation in T_2-weighted MR magnitude images of an US-stimulated bone construct over time: Day 0 (a), After 1 week (b), and after 2 weeks (c)



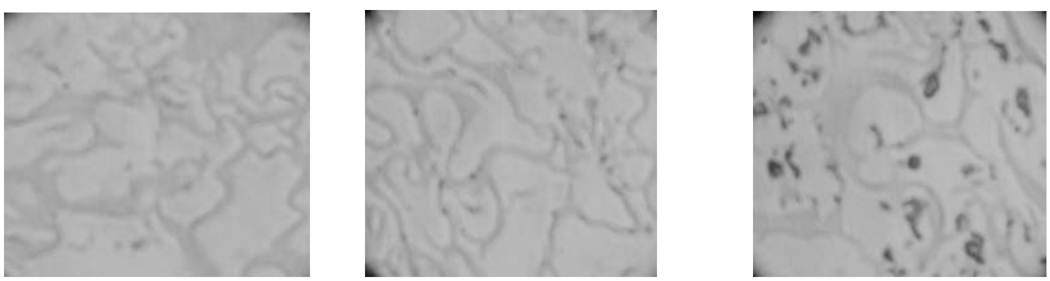
(a) *(b)* *(c)*

Figure 4. Variation in T_2 relaxation time maps of an US-stimulated bone construct over time: (a) Day 0, (b) After 1 week , and (c) After 2 weeks



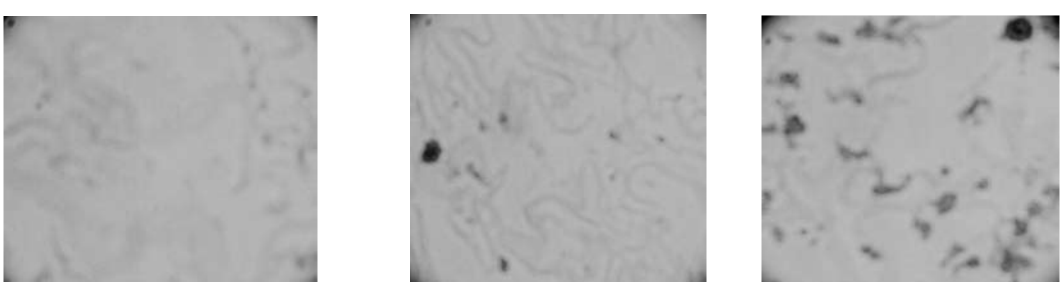
(a) *(b)* *(c)*

Figure 5. von Kossa staining results after 2 weeks of culture for: (a) a control construct, (b) a non-stimulated bone construct, and (c) an US-stimulated bone construct. Image magnification is 20x



(a) *(b)* *(c)*

Figure 6. OCN staining results after 2 weeks of culture for: (a) a control construct, (b) a nonstimulated bone construct, and (c) an US-stimulated bone construct. Image magnification is 20x



(a) *(b)* *(c)*

ultrasound stimulation in accelerating the mineralization of osteogenic constructs *in vitro*; we noted a faster decrease in the T_2 relaxation time values for ultrasound stimulated osteogenic constructs in comparison with the nonstimulated osteogenic constructs.

To our knowledge, this combination of US stimulation and MRM as a monitoring technique has not previously been explored. All the works presented in the preceding section were uniquely exploring either a single aspect of bone growth stimulation or a technique to monitor the growth process (Aaron et al., 2004; Cartmell et al., 2005; Lakes, 2005; Nolte, 2001; Pilla, 2002; Xu et al., 2005). Though our major focus was not the improvement of the tissue engineering methodology by itself (the scaffold features, the biochemical agents, etc.), our work is unique in two aspects. First, it shows that MRM is sensitive enough to characterize the acceleration of osteogenic constructs growth, by sensing the difference in T_2 values between the US stimulated and nonstimulated constructs. Second, it shows that US is effective in accelerating osteogenesis *in vitro*; therefore, our study makes a valuable contribution towards the advancement of bone tissue engineering.

FUTURE TRENDS

Our recent work (Moinnes et al., 2006) made a giant stride in enhancing current bone tissue engineering processes, and also lends support to the conjecture that a combination of stem cell stimulation and noninvasive construct monitoring will increase the efficacy of bone tissue engineering processes. Yet, our work still leaves room for advancement. Notably, the optimized ultrasound parameters have yet to be determined. For instance, the optimal duration or frequency of ultrasound treatment and the critical stage in the cell differentiation process where it would have the greatest effect must be established; knowledge of this type could significantly reduce the net duration of ultrasound administration.

Also, future researchers can study and quantify the effect of various ultrasound operating frequencies on accelerating bone tissue formation in the engineered constructs, and therefore deduce the optimal one for this type of tissue engineering; this would require constructing specifically customized US transducers which would have variable frequencies, since they are not available in the market. Additionally, the optimal MRM parameters for tissue development monitoring must also be determined. These examples of future research directions demonstrate the need for greater knowledge on the optimization of enhanced bone tissue engineering processes.

Another trend which may emerge is the application of conclusions derived from works such as ours in the tissue engineering of other types of biological tissue, such as cartilage and muscle. With regards to the market, a trend which may emerge is that engineered bone tissue may be developed in the manner depicted by us in order to make this type of implant available to patients in a quicker manner, and therefore reduce the waiting periods of patients suffering from bone loss.

Other researchers are currently focusing on different aspects of bone tissue engineering and stimulation methods. For instance, Pinheiro et al. (2006) discussed photoengineering of bone repair processes. Jones et al. (2006) suggested improving the bone tissue engineering process by optimizing bioactive glass scaffolds. Hence, it is through the collaborative efforts of researchers worldwide, who individually and deeply study a given aspect of bone tissue engineering or a certain growth monitoring procedure, that substantial progress in this field may be achieved, the ultimate goal being the ability to provide patients suffering from bone loss with implants that are convenient for them right when they need them.

CONCLUSION

The specific techniques described in this work are not expected to be the ultimate ones that are to be utilized in future bone tissue engineering processes. Several major parameters are to be modified when it comes to the actual application of such techniques in a clinical setting, in order to completely substitute for the traditional autografts and allografts, which are widely employed today. Those parameters are mainly dependent on time and cost, among other constraints. Thus, it is undeniable that we are currently in the very early stages of what is sure to be a long path towards restoring bone tissue in humans using tissue engineering. Much time and research is still needed to progress from these simple experimental preparations, where all environmental and physical conditions can be utterly controlled, to human and clinical settings where unexpected physiological fluctuations introduce uncontrollable variations. Notwithstanding, these experiments are a good starting point on the path to complete bone restoration via tissue engineering.

REFERENCES

Aaron, R.K., Ciombor, D.M., & Simon, B.J. (2004). Treatment of non-unions with electric and electromagnetic fields. *Clinical Orthopaedics, 419*, 21-29.

ASBA Glossary. (2005). Retrieved February 10, 2008, from http://www.azspinabifida.org/gloss.html

Buckwalter, J.A., & Brown, T.D. (2004). Joint injury, repair, and remodeling: Roles in post-traumatic osteoarthritis. *Clinical Orthopaedics and Related Research, 423*, 7-16.

Carnegie Mellon University. (2005). Carnegie Mellon University's Bone Tissue Engineering Initiative. Retrieved February 10, 2008, from http://www.cs.cmu.edu/~tissue/

Cartmell, S.H., Keramane, A., Kirkham, G.R., Verschueren, S.B., Magnay, J.L., El Haj, A.J., et al. (2005). Use of mechanical particles to apply mechanical forces for bone tissue engineering purposes. *Journal of Physics: Conference Series, 17*, 77-80.

Jones, J. Ehrenfried, M., & Hench L.L., (2006). Optimising bioactive glass scaffolds for bone tissue engineering. *Biomaterial*, *27*(7), 964-973.

Kim, H., Suh, H., Sangmee, A., Hyun, W.K., Jung, M.L., Eun, H.K., et al. (2005). In vivo bone formation by human marrow stromal cells in biodegradable scaffolds that release dexamethasone and ascorbate-2-phosphate. *Biochemical and Biophysical Research Communications*, *332,* 1053-1060.

Lakes, R. (2005). *Electrical properties of bone, a review*. Retrieved February 10, 2008, from http://silver.neep.wisc.edu/~lakes/BoneElectr.html

Luo, J., Sun, M.H., Kang, Q., Ying, P., Jiang, W., Luu, H.H., et al. (2005). Gene therapy for bone regeneration. *Current Gene Therapy, 5*(2),167-179.

Mauney, J.R., Volloch, V., & Kaplan, D.L. (2005). Role of adult mesenchymal stem cells in bone tissue-engineering applications: current status and future prospects. *Tissue Engineering*, *11*(5/6), 787-802.

Moinnes, J.J., Vidula, N., Halim, N., & Othman, S.F. (2006). Ultrasound accelerated bone tissue engineering monitored with magnetic resonance microscopy. Engineering in Medicine and Biology Society, 2006. EMBS '06. In *28th Annual International Conference of the IEEE* (pp. 484-488).

Nishimura, D.G. (1996). *Principles of magnetic resonance imaging*. Department of Electrical Engineering, Stanford University.

Nolte, P.A., Klein-Nulend, J., Albers, G.H.R., Marti, R.K., Semeins, C.M., Goei, S.W., et al. (2001). Low-intensity ultrasound stimulates endochondral ossification in vitro. *Journal of Orthopaedic Research, 19*, 301-307.

Nordin, M., & Frankel, V. (2001). *Basic biomechanics of the musculoskeletal system*. Philadelphia, PA: Lippincott Williams & Wilkins.

Pilla, A.A. (2002). Low-intensity electromagnetic and mechanical modulation of bone growth and repair: Are they equivalent? *Journal of Orthopaedic Science, 7*, 420-428.

Pinheiro, A.L., & Gerbei, M.E. (2006). Photoengineering of bone repair processes. *Photomedicine and Laser Surgery, 24*(2), 169-178.

Rahaman, M.N., & Mao, J.J. (2005). Stem cell based composite tissue constructs for regenerative medicine. *Biotechnology Bioengineering, 91*(3), 261-284.

Stock, U.A., & Vacanti, J.P. (2001). Tissue engineering: Current state and prospects. *Annual Review of Medicine*, *52*, 443-451.

Wolff, J. (1986). *The law of bone remodeling*. Berlin; Heidelberg, Germany: Springer-Verlag.

Wu, Y. Shaw, S., Lin, H., Lee, T., & Yang, C. (2006). Bone tissue engineering evaluation based on rat calvaria stromal cells cultured on modified PLGA scaffolds. *Biomaterials, 27*(6), 896-904.

Xu, H., Othman, S., Hong, L., Peptan, I., & Magin, R. (2005). Magnetic resonance microscopy for monitoring osteogenesis in tissue-engineered construct in vitro. Magnetic *Resonance in Medicine*, *54*, 605-615.

Zhang, Y., Cheng, X., Wang, J., Yining, W., Shi, B., Huang, C., et al. (2006). Novel chitosan/collagen scaffold containing transforming growth factor-beta1 DNA for periodontal tissue engineering. *Biochemical and Biophysical Research Communications, 344*(1), 362-369.

KEY TERMS

Bone Differentiation Promoting Factors: Bone differentiation promoting factors are compounds such as dexamethasone, ascorbic acid, and β-glycerophosphate that cause the differentiation of mesenchymal stem cells into osteoblasts.

Bone Tissue Engineering: Bone tissue engineering focuses on the development of bone tissue *in vitro* (using osteoblasts, scaffolds, and bioactive molecules) generally for *in vivo* implantation.

Magnetic Resonance Microscopy: Magnetic Resonance Microscopy is a technique that entails the application of a radiofrequency (RF) pulse to excite a tissue or specimen at resonance, and the acquisition of signal in the form of RF energy during subsequent relaxation of tissue magnetization. This method provides quantitative parameters that are directly dependent on the tissue properties.

Mesenchymal Stem Cells: Mesenchymal stem cells are multipotent stem cells capable of differentiating into cells such as osteoblasts, adipocytes, chrondroblasts, fibroblasts, and muscle cells.

Osteoblasts: Osteoblasts (bone-forming cells) produce osteoid, a component of the bone matrix, and cause mineralization of the matrix.

Scaffold: In bone tissue engineering, a scaffold is a biomaterial surface used for the development of new bone tissue *in vitro*.

Ultrasound Stimulation: Ultrasound waves are sound waves with frequencies greater than 20,000 Hz. Low intensity pulsed ultrasound has been found to be effective at stimulating the differentiation of mesenchymal stem cells into osteoblasts.

Agent-Based Patient Scheduling

Einar Jakob Hovland
Agder University College, Norway

Even Åby Larsen
Agder University College, Norway

INTRODUCTION

Agent-oriented software engineering has, by many researchers, been dubbed the new paradigm in software development, and from its original concepts in the early '80s, agents and agent systems are now active research areas in computer science.

This evolution offers a promising approach to the development of patient scheduling systems. Coordinating and processing a vast amount of complex variables, such a system should be designed to stock and schedule a wide range of resources based on the patients' health condition and availability, drawing on the advantageous data control and optimization abilities of agent technologies.

This article presents the design of a working agent-based patient scheduling system prototype.

BACKGROUND

Agent Systems Definitions

A software agent is an autonomous entity capable of performing actions and interactions typically based on notions of beliefs and goals. In addition to autonomy and pro-activeness (Wooldrige & Ciancarini, 2001), typical characteristics of agents are anthropomorphism, situatedness, and social ability.

Agent systems can consist of just one such agent or a collection of agents performing different tasks based on individual or common goals. Due to the individualistic characteristics described above, an agent system can collectively draw on further advantages, including mobility, dynamic sizing, and complex cooperation through negotiation.

Historical Context

The term agent can be traced back to the *Actor Model* first presented by Hewitt, Bishop and Steiger (1973). This early concept simply defined agents as entities with a memory address and computational behavior to help solve common tasks.

In the late '80s, the Belief-Desire-Intention (*BDI*) model was proposed (Bratman, Israel, & Pollack, 1988). The model represented a novel approach of giving human properties to digital agents. Through available information about the environment (beliefs), the agents are given a set of certain possible actions (desires) which are activated based on agent goals (intentions).

With sophisticated communication, agents can interact to cooperatively achieve global tasks and goals. But this coordination needs more than sufficient shared semantics. It also requires planning and scheduling techniques to govern the order and partition of tasks. Roughly, there are two general frameworks developed over the last decades to deal with these challenges; namely, the partial global planning (PGP) algorithms, and the joint intentions framework.

PGP (Durfee & Lesser, 1991) was an early attempt at planning in a distributed dynamic environment. By sharing and communicating intentions globally, the framework allowed agents to make optimal decisions locally.

Joint intentions present another approach in coordinating and planning node actions. Instead of passively collecting information to decide on optimal actions, the joint intentions frameworks is mainly focused on reaching common goals through agreement (Jennings, Sycara, & Wooldridge, 1998). The model is that of a team's intention, rather than the individual agent's goals. As the main focus of communication in these frameworks is reaching agreements, a natural consequence is *negotiation*.

Almost all negotiation in agent systems is based on some notion of auction (e.g., Luck, McBurney, Shehory, & Willmott, 2004). The bids used in auctions are often based on game theory mechanics and utility functions. Basically, game theory is the study of decisions in environments where several players interact (Vlassis, 2003). Game theory frameworks include the Nash Equilibrium solution concept (Nash, 1950) and Operations Research (Phillips, Ravindran, & Solberg, 1976), amongst others.

Application to Patient Scheduling Systems

Decision support systems and patient scheduling systems in particular, have become an increasingly important factor in many hospitals and medical institutions (Manansang & Helm, 1996). The primary goal of patient scheduling systems is to treat as many patients as possible in the shortest possible time (Bartelt, Lamersdorf, Paulussen, & Heinzl, 2002).

The examination and treatment process for patients involves a high degree of uncertainty regarding time spans and the resulting diagnosis, thus patient scheduling systems have been deemed complex (ibid.). Modern patient scheduling system design focuses on patients, rather than specific tasks or resources (Guo et al., 2004).

Hence, patient scheduling systems exhibits many of the same characteristics as those recognized in the agents and agent systems literature. Characteristics like entity-focused design, and high complexity and abstraction levels are well-founded identifiers in agent-oriented literature

Some earlier proposals exist—most noteworthy, the MedPage project (e.g., Bartelt, Wagner, & West, 2002), which is an ongoing attempt to introduce agent planning and scheduling systems at German hospitals.

AGENT SCHEDULING SYSTEM DESIGN

Following the theories presented, this article proposes a patient scheduling technique founded on software agents. Using well-established optimization theories from various fields of science, including optimal decision processing and game theory, this section will present an agent system labeled *AgentMedic* to effectively schedule patients in a medical institution.

The following four subsections will describe this system in detail. First, we define the three distinct agent types used in the system, focusing on their tasks and goals. Second, the communication and utility data flow between these agents are introduced. The third subsection defines the optimal decision functions used to determine the relevant value of patients and their position in the treatment cycle. And lastly, the optimization and scheduling processes are presented.

Agent Specifications

When choosing the types of agents needed in such a system, it is convenient to remember the typical agent characteristics presented earlier in this article. The design should allow for the agents to make use of their anthropomorphic and pro-active nature, so as to represent a live entity in the best possible way (Foner, 1993). Furthermore, autonomous entities requiring both social and flexible behavior must, in particular, be considered for agent abstraction (Wooldridge & Ciancarini, 2001).

This in mind, *AgentMedic* contains three distinct types of agents, each representing different roles in the context of a treatment process and the actions performed in a medical institution:

- A patient agent is responsible for coordinating the full examination and treatment process of the patient it is representing. The agent must, at all times, keep track of the progress of the patient, his/her health condition, and the management of upcoming examination or treatment tasks.
  - **Goals:** *Get examined/treated as fast as possible*
  - **Possible actions:** *Apply for appointment*
- A personnel agent represents a member of the hospital staff. Its tasks involve keeping track of the personnel competence as well as availability. Further, a personnel agent must use these properties to optimize and coordinate examination and treatment tasks with the patient agents.
  - **Goals:** *Treat/examine patients as fast as possible*
  - **Possible actions:** *Grant/deny appointment, examine/treat patients.*

- An equipment agent is, in the same manner as the personnel agents, responsible for keeping track and coordinating the usefulness and availability of hospital equipment. The agent is involved in optimizing and coordinating the assignment of equipment to treatment tasks, taking patient and personnel into account.
  - **Goals:** *Support treat/examination process effectively*
  - **Possible actions:** *Grant/deny participation, support treat/examination process of patient.*

Communication Patterns and Utilities

The figure below show the general communication patterns for the agents defined above. A patient requests an examination or treatment within some field of expertise. The information from all patients is then dealt with by the equipment and personnel agents to reach the optimal scheduling for a particular timeslot cycle. The results are sent back to the patient agents who at the next timeslot can move on to next step in the treatment cycle (Figure 1).

To fulfill the optimization potential of an agent system, every agent—whether it represents a patient, a representative from the personnel, or a piece of equipment—constantly needs to argue their position and usage in the system. For example, a patient argues for urgent treatment or examination based on his or her health condition, a doctor argues for using his or her competence, and equipment argues for being used in suitable tasks.

To represent values for this type of argumentation, we introduce the notion of utility. A utility value is a comparable measure, which in our case is used to weight the health condition of patients, the competence of personnel and the usefulness of equipment. In addition, and in a similar manner, *AgentMedic* use an availability value (either 1 for available or 0 for not available) to represent the availability of personnel and equipment. This value denotes whether a piece of equipment or a member of personnel is present at the institution.

Patient Agent Task Generation

Assuming we know the health condition of patients and the probability of possible diagnostic results, we can introduce a simple example using two patients: John and Mary. For simplicity, only two patients are considered, and they both undergo exactly three cycles in a treatment process: a preliminary examination, a secondary examination, and a treatment process.

Figure 1. Communication patterns amongst agents

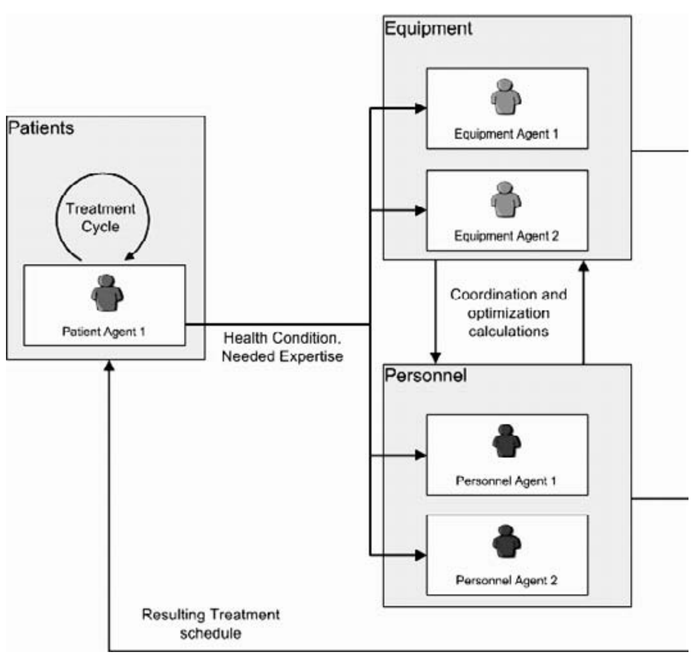
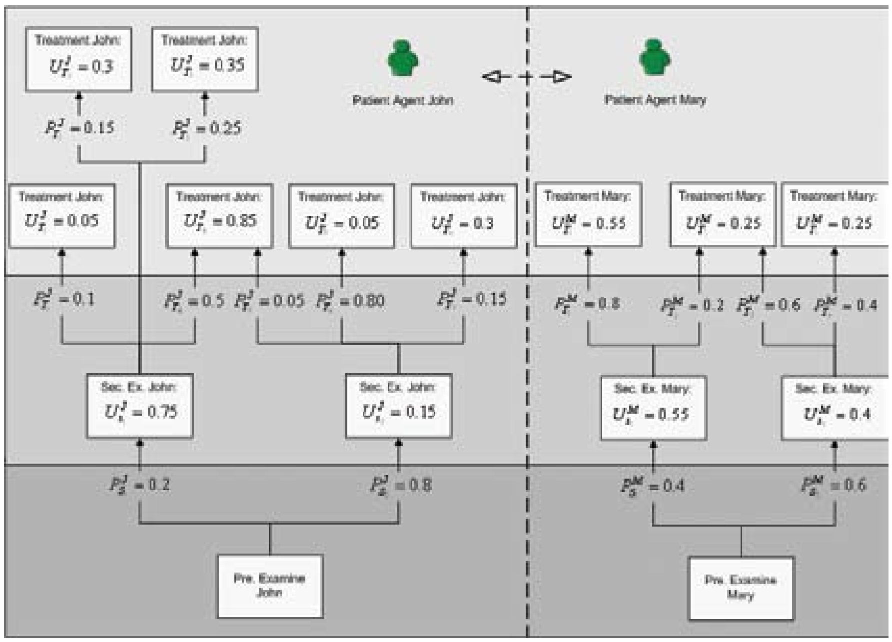


Figure 2. John and Mary parallel patient agent treatment processes



Combining all three stages and including example utilities *U* and probabilities *P*, the two patient agents can be presented in Figure 2.

The figure above shows the process path tree of two patient agents: one representing John, and one representing Mary. Assuming there is only one doctor available, the two agents should cooperate and bargain to obtain effective scheduling for the different stages of the two patients.

The scheduling is decided by using the relevant probability *P* and utility *U* values at each stage. By knowing the utility value *U* and the probability *P* of reaching all achievable states *s*, an optimal action α_t^* can be calculated by using the Bellman equation:

$$\pi^*(s) = \arg\max_{\alpha} \sum_{s^1} P(s^1 \mid s, \alpha) U^*(s^1) \qquad \text{(i)}$$

In words, for each state a related utility is calculated by finding the max return value of all possible actions obtainable in this state. The max return value is the utility multiplied by the probability for each new obtainable state. In our medical institution context we need to find these temporary utilities to be able to prioritize one task over another so that resources like personnel and equipment can be used effectively.

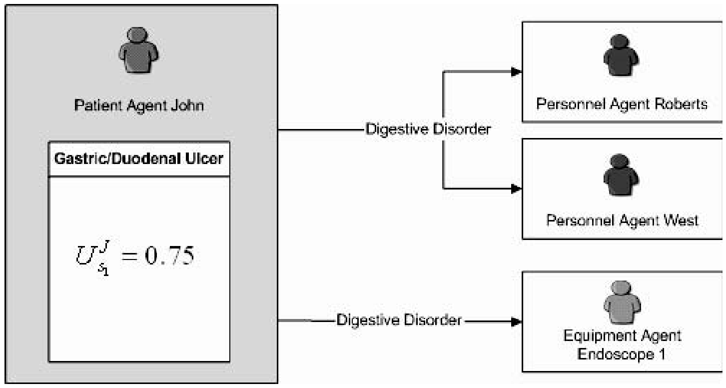
Using our example from Figure 2, the two agents need to agree on which patient goes first for their preliminary examination. Using the formula in (i), we can calculate the total utility values for John and Mary:

$$\alpha_{pre.ex.John} = (0.2 * 0.75) + (0.8 * 0.15) = 0.27$$
$$\alpha_{pre.ex.Mary} = (0.4 * 0.55) + (0.6 * 0.4) = 0.46$$

Hence, the results indicate ($0.27 < 0.46$) that preliminary examination of patient Mary should be prioritized before preliminary examination of John. Again, this is based on the possible outcomes of the examination. After the preliminary examination, we assume Mary was diagnosed with migraines ($U = 0.4$) and her agent now has to calculate the new temporary utility and argue for being prioritized at the next stage. By applying this formula for each patient on each stage, task priorities can easily be measured, compared, and prioritized.

As explained in the communication section above, the patient agents distribute the utilities to personnel and equipment agents for evaluation. When a utility is distributed, we may consider this action as a task generation, proposing a new task for the personnel and equipment agents. Naturally, many such task generations may occur within the same timeslot, generating the need for comparing and prioritizing tasks. The next section will take a closer look at this process.

Figure 3. Example: John's role assignment



Optimization and Scheduling

Before personnel and equipment agents can be assigned to the various treatment tasks presented by the patient agents, we need to investigate their competence and usefulness utilities in order to filter out unrelated personnel and equipment from our scheduling process equations.

This is done by relating the competence and usefulness of personnel and equipment to the general medical field of the task at hand. As an example, the symptom *heartburn* and the condition *ulcer* are both assigned to the more general condition of *digestive disorder*. Further, digestive disorder is part of the competence field for some of the personnel agents and has a usefulness factor in some of the equipment agents. Hence, the tasks involving the symptom heartburn or the condition ulcer are linked with personnel having competence in—and equipment with—a usefulness factor for digestive disorders.

The result of this process is patient agent tasks connected to "wish lists" containing desirable personnel and equipment.

Figure 3 shows patient agent John's resource wish list for the upcoming examination task. To decide the allocation of available personnel and equipment to patients, *AgentMedic* uses concepts from game theory, and more specifically the Nash Equilibrium (NE) model. The NE formula allows a consideration of all available agents and their respective utility and availability values at once, ensuring effective optimization:

$$u_i\left(\alpha_{-i}^{*}, \alpha_i^{*}\right) \geq u_i\left(\alpha_{-i}^{*}, \alpha_i\right) \qquad \text{(ii)}$$

In words, the formula state that a *joint action* (for example, "John treated by Smith" and "Smith treats John") is NE if no agent can improve utility by choosing a different action. Thus, NE represents optimal distribution for the two agents involved.

For example, consider a timeslot generating four tasks by four different patient agents: John, Mary, Chris, and Sarah. As shown, the four agents define medical field for their tasks and add personnel to their list in Figure 4.

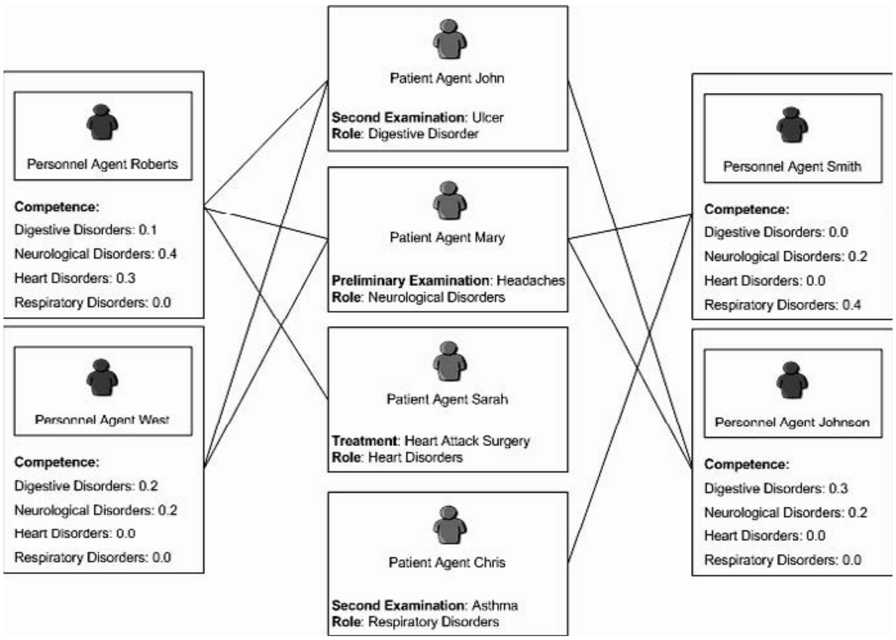
The diagram above shows that each patient agent finds all personnel agents that have some competence within the role that is assigned to their next task. For example, John would, in this example, make a sorted list with the following personnel agents: {Johnson, West, Roberts}.

Now that related tasks and personnel have been tied together, we can use a matrix to find NE. The matrix shows both the competence utility and a relative availability factor:

	Roberts	**West**	**Smith**	**Johnson**
John	*0.1 , 0.0*	*0.2 , 0.85*	*0.0 , 0.85*	*0.3 , 0.85*
Mary	*0.4 , 0.3*	*0.2 , 0.3*	*0.2 , 0.3*	*0.2 , 0.0*
Sarah	*0.3 , 0.15*	*0.0 , 0.15*	*0.1 , 0.15*	*0.0 , 0.0*
Chris	*0.0 , 0.5*	*0.0 , 0.5*	*0.4 , 0.5*	*0.0 , 0.0*

Note that the availability factor (1 for available or 0 for not available) has been multiplied with the patient utility. The three factors (utility, competence, and availability) are now part of the same matrix, and can be calculated using the NE formula. To calculate optimal distribution, we apply (ii) to each matrix cell.

Figure 4. General role assignment example



This produces the following set of NE joint actions:

{(John, Johnson), (Chris, Smith), (Mary, Roberts)}

With the correct personnel assigned, the patient agents should connect to the appropriate equipment agents in the exact same manner. Only patients that have been assigned a personnel representative may look for a piece of equipment however, as with no personnel, there will be no examination or treatment. We find the following results when applying the NE formula to each joint action:

{(John, Endoscope), (Mary, Surgery Equipment)}

Hence, all three patients but Chris were assigned relevant and available equipment for their task. As we can see, the NE does not always give fully satisfactory results. This is due to having two stages of NE, first for personnel then for equipment. As no equipment was available for Chris, we need to check if our last patient, Sarah, can be assigned to the now-available personnel Smith. As this can be the case for several patients, we run the NE matrix again with the remaining patients and personnel:

{(Sarah, Smith)}

Finally, we run the last NE equation, checking if the necessary equipment for Sarah is available. We achieve the following result:

{(Sarah, Blood Test Set)}

As a result of this, patients John, Mary, and Sarah are scheduled for examination/treatment in this timeslot. Chris will have to wait for the next timeslot where the necessary equipment hopefully is available.

Prototype Implementation

The ideas above have been implemented in prototype form to investigate the feasibility of the concepts and to verify that such a system will possess the robustness and flexibility claimed to be properties of agent-based systems (Hovland, 2006).

We have simulated the dynamic nature of hospitals by adding and removing agents from a running system. This represents changes in staff, equipment, and patients. We have also simulated mobility between institutions by moving agents from one running system to another. The tests have been mostly successful. The problems encountered can be attributed to shortcomings in the prototype, rather than the concept itself (ibid.).

FUTURE TRENDS

In accordance with the evolution of information systems in general, modern, health-specific systems have increased in both size and complexity. Typical modernizing trends within this field include globalization of system structure, emerging technologies, new complex data types, ubiquitous computing, and inclusion of strategic and administrative data (Haux, 2005).

Patient scheduling systems being part of this step to the future require, at the same time, a sharp focus on the patient (Guo et al., 2004). We have seen that agent technologies represent a promising approach in dealing with both the increase of complexity, while keeping this focus at bay.

CONCLUSION

Agent-based systems is an emerging approach in dealing with modern complex information systems. Furthermore, agents are well suited to deal with large amounts of variables, using negotiation to deal with complex planning procedures. As such, this article presents an agent-based patient scheduling system based on planning and optimization algorithms from game theory.

While the developed prototype from Hovland (2006) shows that such an approach is feasible, it is still a long way from real-life application. The presented approach demonstrates the core scheduling operations of such a system, but assumes some simplifications especially in regards of utility functions of the patients.

REFERENCES

Bartelt, A., Lamersdorf, W., Paulussen, T. O., & Heinzl, A. (2002, May). Agent oriented specification for patient-scheduling systems in hospitals. In *Proceedings of IFBI, Dagstuhl*. Saarland, Germany.

Bratman, M. E., Israel, D. J., & Pollack, M. E. (1988). Plans and resource-bounded practical reasoning. *Computational Intelligence*, 4(4), 349–355.

Durfee, E. H., & Lesser, V. R. (1991, September). Partial global planning: A coordination framework for distributed hypothesis formation. *IEEE Transactions on Systems, Management and Cybernetics, 21*(5), 1167–1183.

Foner, L. (1993, May). *What's an agent, anyway? A sociological case study*. Agents Memo 93-01, M.I.T. Media Lab. Massachusetts Institute of Technology, Cambridge, MA.

Guo, M., Wagner, M., & West, C. E. (2004). Outpatient clinic scheduling – a simulation approach. In R. D. Ingalls, M. D. Rossetti, J. S. Smith, & B. A. Peters (Eds.), *Proceedings of the 2004 Winter Simulation Conference* (pp. 1981–1987).

Hewitt, C., Bishop, P., & Steiger, A. (1973, August). A universal modular actor formalism for artificial intelligence. *International Joint Conferences on Artificial Intelligence* (Vol. 3, pp. 235–245). International Artificial Intelligence Center, Menlo Park, CA.

Hovland, E. J. (2006, December). *Positive and negative factors in agent oriented software development—A case study*. Unpublished master's thesis in Information Systems, Faculty of Economics and Social Sciences, Agder University College.

Jennings, N. R., Sycara, K., & Wooldridge, M. (1998). A roadmap of agent research and development. *Autonomous Agents and Multi-Agent Systems, 1*(1), 7–38.

Jennings, N. R., & Wooldridge, M. (2000). *Agent-oriented software engineering*. In J. Bradshaw (Ed.), *Handbook of Agent Technology*. AAAI/MIT Press.

Luck, M., McBurney, P., Shehory, O., & Willmott, S. (2005). *Agent technology roadmap: A roadmap for agent based computing.* AgentLink, Electronics and Computer Science. University of Southampton, United Kingdom. Retrieved February 7, 2008, from http://eprints.ecs.soton.ac.uk/11788/01/al3roadmap.pdf

Manansang, H., & Helm, J. (1996). An online simulation-based patient scheduling system. In *Proceedings of the 1996 Winter Simulation Conference.*

Nash, W. (1950). Equilibrium points in n-person games. In *Proceedings of the National Academy of Sciences of the United States of America* (Vol. 36, pp. 48–49).

Phillips, D. T., Ravindran, A., & Solberg, J. (1976). *Operations research: Principles and practice*. Canada: John Wiley & Sons.

Vlassis, N. (2003, September). *A concise introduction to multiagent systems and distributed AI*. Informatics Institute, University of Amsterdam, Amsterdam, the Netherlands. Retrieved February 7, 2008, from http://staff.science.uva.nl/~vlassis/cimasdai/cimasdai.pdf

Weyns, D., Steegmans, E., & Holvoet, T. (2004). Towards active perception in situated multi-agent systems. *Journal on Applied Artificial Intelligence, 18*, 9–10.

Wooldridge, M., & Ciancarini, P. (2001). Agent-oriented software engineering: The state of the art. In P. Ciancarini, & M. Wooldridge (Eds.), *Agent-Oriented Software Engineering* (Vol. 1957). Springer-Verlag Lecture Notes in AI.

Zambonelli, F., Jennings, N., & Wooldridge, M. (2003). Developing multiagent systems: The gaia methodology. *ACM Transactions on Software Engineering and Methodology, 12*(3), 317–370.

KEY TERMS

Agent: "A computer system that is capable of flexible autonomous action in dynamic, unpredictable, typically multi-agent domains" (Luck et al., 2005).

Agent System: "A group of agents that can potentially interact with each other" (Vlassis, 2003). "In multi agent systems, applications are designed and developed in terms of autonomous software entities (agents) that can flexibly achieve their objectives by interacting with one another in terms of high-level protocols and languages" (Zambonelli, Jennings, & Wooldridge, 2003).

Anthropomorphism: The attribution of human characteristics to software agents, including personalizability, learning and reasoning (Foner, 1993).

Dynamic Environment: Agent systems allow agents to enter, leave and change at runtime.

Mobility: The ease of transferring agents to other systems with the same characteristics, due to their autonomous nature and the dynamic property of agent systems.

Negotiation: Often used amongst agents as a means of optimal decision making. Agents can exhibit the ability to perform an optimal individual action, which contributes to the goals of the system as a whole.

Situatedness: Agents are in close relationship to its environment. Agents optimally have awareness capabilities, supervising their domain through sensors and performing reactive actions through effectors accordingly (Jennings & Wooldridge, 2000; Weyns, Steegmans, & Holvoet, 2004).

Social Ability: Agents communicate and reach mutual decisions through conversation and negotiation (Foner, 1993; Wooldrige & Ciancarini, 2001).

Analytics: Unpacking AIDS Policy in the Black Community

Fay Cobb Payton
North Carolina State University, USA

INTRODUCTION

With advances in information technology analytics applications, society stands to benefit greatly from health care innovation. The ability to link physicians, hospitals, pharmacies, clinics, and patients to health information networks and clinical and financial data management and analyses can prove to be invaluable in the diagnosis and treatment of chronic episodes of illnesses such as AIDS/HIV.

This access to data is a necessity in order for hospitals and physicians to provide the highest level of safety and quality of care. Providing the correct diagnosis and procedures is critical for the patient's utmost care. With the high costs associated with AIDS/HIV procedures, medications, and physician consultants, the integration of IT can offset these costs and improve the efficiency of the organizations. Factors such as cost of care and length of stay continue to drive health service delivery, resource availability, and quality of care.

Business analytics (BA), often termed business intelligence (BI), applications can carefully provide insight into the (in)significance of these factors in health care systems' abilities to treat AIDS/HIV in general. In particular, demographic variables that relate to cultural, socioeconomic status and community dimensions of those most impacted (namely, Black Americans in the United States, which is the focus of this writing) by the AIDS/HIV epidemic are often disregarded. For the broader community, the questions to address are diverse. What can business analytics inform us about Black Americans infected by AIDS/HIV? What are the broader cultural issues that often are not modeled by analytical tools? How do these findings stand to impact public policy and how the health care community can better assist those living with the disease? In this chapter, I take on these questions by first reviewing major issues and trends in AIDS/HIV and IT literatures by focusing on health disparities in one historically underserved group; namely, Black Americans. Next, I present public health conceptual framework that augments this discourse by depicting those factors uncovered in traditional information technology/systems works. This chapter concludes with recommendations for future research opportunities for examining AIDS/HIV public policy issues.

BACKGROUND

According to *Data Bulletin* (2003), between 1991 and 2003, per capita spending on health care in the United States rose almost 95%, with little improvement in national health metrics. Among policymakers, well-regarded media outlets, and others (Kovak, 2005), there is widespread disagreement about a final solution to the problem of rising health care costs. Moreover, there is equally widespread agreement that one element must be a large-scale, systemic change in the uses of information technology for health care management and delivery.

Comprehensive IT systems have improved efficiency and productivity in virtually every major industry, with the conspicuous exception of health care, based on recent RAND reports (Fonkych & Taylor, 2005). Used primarily for administrative tasks such as billing and scheduling, IT offers great promise for use in Electronic Medical Record Systems (EMR-S) or as a clinical diagnostic aid.

The AIDS/HIV epidemic continues to have a riveting impact on the United States. In order to slow the epidemic, analytics enables the field to improve upon its understanding of the dynamics behind the disease. There are an estimated 800,000 to 900,000 people currently living with AIDS/HIV in the United States, with approximately 40,000 new AIDS/HIV infections occurring in the United States every year. More recently, gender has become a significant factor to pay attention to when identifying new cases each year. For several years, men dominated the estimates of new infections; women, in general, are now also significantly affected, and Black women, in particular. Adopted from the

Centers for Disease Control and Prevention (CDC), Figure 1 shows that 70% of new HIV infections each year occur among men, although women are also significantly affected and hold the other 30%.

With regard to race, more than half of newly affected AIDS/HIV cases occur among Black Americans, although this subpopulation only represents 13% of the total population of the United States. Hispanics, who make up about 12% of the United States population, are also disproportionately affected. Figure 2 shows the estimates of annual new infections according to race. Blacks lead with more than 50%, while White and Hispanics trail behind with significantly lower percentages.

An even more astounding statistic shows that Black women accounted for 72% of all new diagnosed AIDS/HIV cases in the United States. The AIDS diagnosis rate for Black men was nearly 11 times greater than their white male peers, and this rate is 23 times greater for Black women in comparison to white women (CDC, 2005; LaVeist, 2005).

Business analytics (BA) focuses on effective use of these data and information to drive positive business actions such as those already noted. The body of knowledge for this area includes both business and technical topics, including concepts of performance management; definition and delivery of business metrics; data visualization; and deployment and use of technology solutions such as OLAP, dashboards, scorecards, analytic applications, and data mining (http://www.tdwi.org/). Analytics technologies reduce uncertainty, predict with precision, and optimize performance, and can be inclusive of forecasting, text and data mining, and statistical methods, just to name a few. These technologies can enable health care organizations to effectively use:

- **Electronic medical records (EMR).** Stores all patient information in a centralized location and allows physicians, nurses, patients, and other clinical staff to access clinical and financial data spontaneously.

Figure 1. Estimate of annual new infections by gender (N = 40,000)

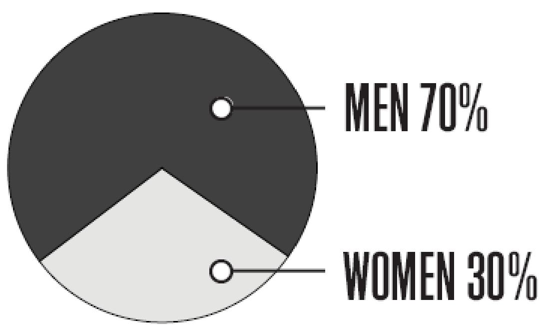
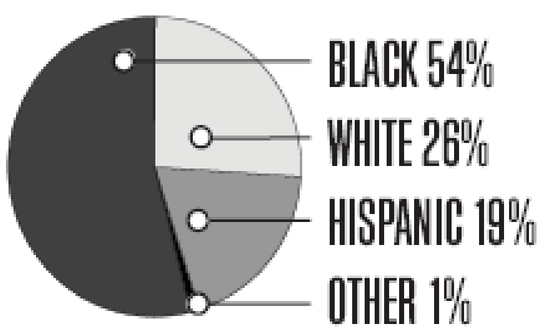


Figure 2. Estimate of annual new infections by race



- **Computerized physician order entry systems (CPOE).** Manages order and fulfillment of medical actions, including prescriptions, lab orders, discharges, transfers, and referrals.
- **Clinical decision support systems (CDSS).** With the increased use of evidence-based medicine and the need for currency among clinical professionals, innovative knowledge can impact patient outcomes. These systems provide information on diagnosis, treatments, and recommendations given the situation.
- **Wireless networks and remote care.** This technology allows valuable information and observation beyond a hospital bed.
- **Aggregate database on treatment outcomes.** Placing data in databases allows faster assessments and analysis of the data. The data can be quickly filtered to show procedures, treatments, medications, and costs. Valuable information can be pulled much more quickly from electronic sources than hard copies.

BUSINESS ANALYTICS: HOW IT (UN)INFORMS US

Business analytics (BA) can better inform hospitals, insurers, and providers of two critical factors in the wake of managed care and capitation: cost of care (COC) and length of stay (LOS) for HIV/AIDS cases. Factors such as a patient's age, race, gender, and high-risk behaviors are significantly influential in determining COC and LOS. Moreover, geographic location has emerged as a critical factor in the equation. Recent reports by Newsweek (2006), UNAIDS (2004), and the Centers for Disease Control and Prevention (2005) confirm the notion that HIV/AIDS cases have spiked disproportionately in rural United States (particularly the South) and sub-Saharan African countries (particularly Kenya, Tanzania, and South Africa) where people of color comprise significant proportions of the total population.

Using stepwise regression models, SAS Enterprise Guide software (see www.sas.com) and a patient de-identifiable dataset of nearly 90,000 hospital encounters, the author sought to determine how many African American persons had HIV/AIDS. Table 1 shows that roughly 75% of the cases in the dataset represented Blacks, while 23% were whites. Figure 3 plots these figures along the normal curve to illustrate the vast differences in the number of cases between the two groups.

Within the framework of business analytics, these findings, along with other relevant statistics, can be used to inform performance management applications. Within a health care domain, these applications can measure, monitor, and evaluate performance relative to physician, nursing, and other clinical staff retention, internal operations, and clinical care; process performance is relative to disease intervention and diagnostic related groups (DRGs) and productivity enhancement. Hence, these data enable health care organizations to fuse typically disparate clinical and financial data and result in evidence-based medicine/practice.

Balanced scorecards can be generated with these objectives: decrease cost of care, decrease length of stay, and reduced averages by 10% among HIV/AIDS cases in general and Black Americans in particular. Figure 4 shows a snapshot from the SAS Strategic Performance Management (SPM) application to help health care providers meet these objectives. SPM offers graphical representations with diagrams, color-coding, and key

Table 1. Frequency of HIV/AIDS cases found in medical dataset

Race (uniform)				
RACE	**Frequency**	**Percentage**	**Cumulative Frequency**	**Cumulative Percentage**
1	375	23.99	375	23.99
2	1175	75.18	1550	99.17
3	6	0.38	1556	99.55
5	4	0.26	1560	99.81
6	3	0.19	1563	100.00

performance indicators (KPIs) functionality to aid in and enhance decision-making.

While BA applications and methods inform of valued statistics in the plight against AIDS/HIV, the numbers do not tell the full story. This is particularly the case in the Black community. In its May 2006 article, "Does Class Trump Race? DiversityInc suggested that to understand health outcomes and disparities, one must not ignore socioeconomic status (SES). This translated into those who have the highest propensity to be uninsured or underinsured; namely, Blacks, Latinos, and those living in impoverished rural areas. Hence, according to Isaacs' article in the *New England Journal of Medicine* (2004), many of those living in lower SESs are Black or other ethnic minorities.

Further, much of the health care system is based largely on one's educational training and the ability to navigate treatment, service delivery, and a host of payer guidelines and policies, all of which warrant some degree of adequate financial and educational resources. Oftentimes, this is not the case of underrepresented groups who are infected and affected by AIDS/HIV in the Black community. Health care policy experts (LaVeist, 2005; Shi & Stevens, 2005) liken these outcomes to race/ethnicity disparities where "adequate measures, such as cultural factors and measures of discrimination have not been developed or implemented, so we are left with race/ethnicity measures serving as relatively inaccurate proxies" (p 35).

Much of the demographics captured in the 90,000 records dataset extends beyond the scorecard in Figure 4 and includes ICD-9 codes, total hospital charges incurred by the patient, diagnoses (as many as 16 per record), medical record numbers, and attending physician codes. While BA applications can stimulate increased comprehension of the correlations among these variables to improve process improvement, forecasting, causal modeling, and resource utilization, just to name a few, clinical and financial systems often do not capture community determinants of vulnerability among chronic disease episodes as classified by Aday (1994). These vulnerable diseases or health conditions link to both the community and individual levels of resources. The community resources encompass cohesiveness among people and neighborhood characteristics (e.g., presence/absence of violence, unemployment rates, access to physical recreation) while the individual level encompasses social and human capital, social status, and health needs. Aday (1994) suggests that there are nine subpopulations that comprise vulnerability. Persons living with AIDS/HIV fall into one of these vulnerable subgroups.

Moreover, Stevens and Shi (2003) offer a conceptual model linking race and ethnicity with health care experiences. Their model encapsulates family characteristics that denote societal and environmental drivers and health system characteristics capturing health and social policy. While socioeconomic status or income

Figure 3. Normal curve showing Whites (case 1) and Blacks (case 2)

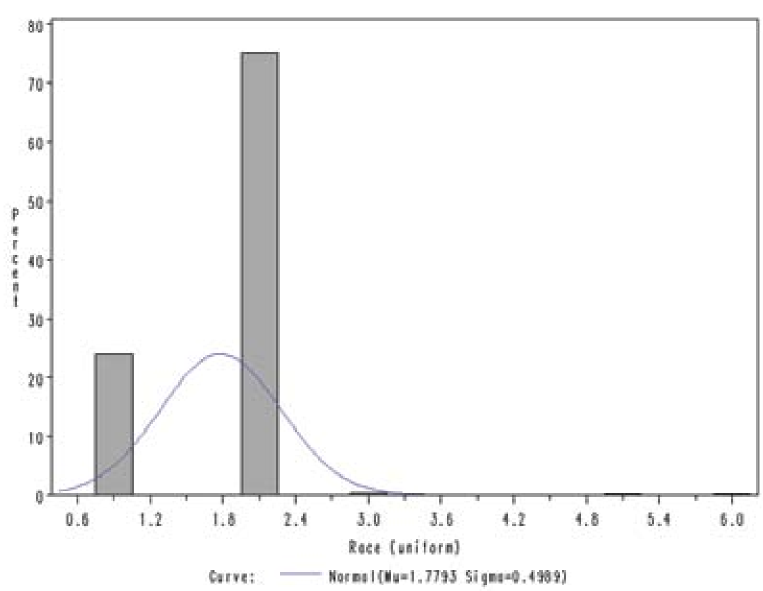
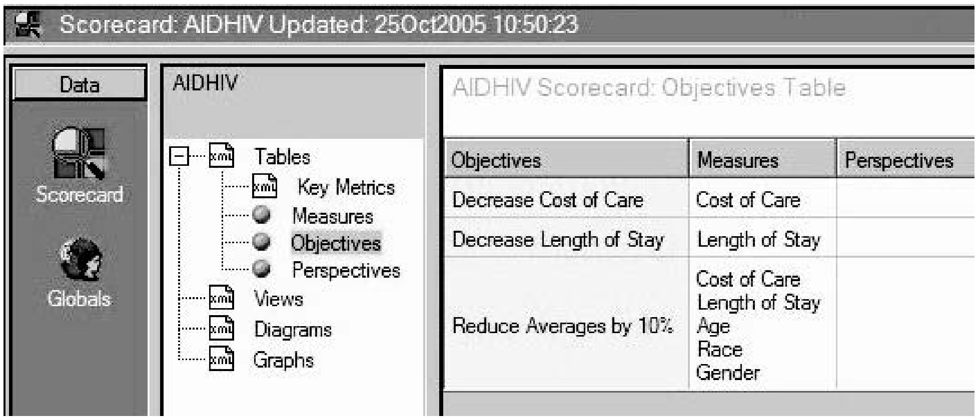


Figure 4. Example of scorecard used to minimize cost of care and length of stay, and reducing these averages among AIDS/HIV cases



can trump race, as suggested by the Kim (2006) article, other dimensions of discourse cannot be ignored; namely, cultural factors, discrimination, health needs, provider factors, and health system factors. Cultural factors rest largely on family structure norms or beliefs and language that impact care delivery, behaviors, and health decision-making and seeking. Discrimination provides the pathway to link race to (un)intended motives, practices, and care delivery, thereby resulting in negative consequences associate with the medical system in general. Health needs influence services sought and implementation of recommendations by patients. Provider factors examine how patients interact with physicians and vice versa. It has been well documented that "physician perceptions and beliefs about patients are affected by race, ethnicity and socioeconomic status" (Shi & Stevens, 2005, p. 43) and do impact physicians' delivery of services and treatment plans. Lastly, health system factors examine the availability of health care resources and insurance coverage.

Furthermore, each of these dimensions can be linked to health disparities, particularly among Black Americans living with AIDS/HIV and dwelling in vulnerable scenarios of low social and human capital and high unemployment. Among the most vulnerable are those residing in the rural South (USA) where AIDS/HIV inflection rates are increasing at alarming rates while literary rates are often less than 30% for completing a high school education. These areas are characterized by the presence of high unemployment rates, incarceration among Black men, drug trafficking, and migrant workers.

FUTURE TRENDS AND CONCLUSION

Business analytics are applications and methodologies that can augment our understanding of a phenomenon in question. In this case, BA was used to analyze a de-identifiable set related to medical encounters or visits to a teaching hospital. BA offers statistical analyses, forecasting, and scorecards to describe the AIDS/HIV cases discovered in the dataset. However, BA is only as sound as the data provided to its users. Oftentimes, this will require that health care researchers analyze more than one dataset or examine findings from other reference disciplines to enrich their interpretation of results. The goal here is to offer the reader the opportunity to step back and reflect on the discourse at hand and recognize that chronic diseases such as AIDS/HIV in vulnerable populations (e.g., Blacks in rural South or in significantly vulnerable scenarios) are not monolithic. A deeper understanding of broader public health frameworks can shed auxiliary viewpoints on the data, thereby affecting one's interpretation of the results.

Hence, critical to this dialogue is the awareness of health literacy, cultural competency, and health communications in this domain. While there are numerous Internet resources that cover Black Americans' health, disparities, and AIDS/HIV, the current system is best exemplified by the absence of adaptive structuration (DeSanctis & Poole, 1994). That is, unintended community structures and human action can and often do emerge under the conditions of advanced technologies' designs where IT designers and sponsors are not void of predefined notions of structuration. Health care Web sites can lack the presence of cultural competency and a substantial degree of effectiveness in disseminating AIDS/HIV and other pertinent medical information to the Black community. To this end, Healthy People 2010 (2006) concluded:

Often people with the greatest health burdens have the least access to information, communication technologies, healthcare, and supporting social services. Even the most carefully designed health communication programs will have limited impact if underserved communities lack access to crucial health professionals, services, and communication channels that are part of a health improvement project.

Lastly, how AIDS/HIV issues are communicated to targeted groups could augment findings from clinical and financial applications that serve as input into BA applications. Health communication scholars argue that the media do not simply report on health issues, but rather play a major role in setting the agenda of health issues for the general public (Pitts & Jackson, 1993). This school of thought agrees with scholars that have analyzed the influence of the media in setting the public agenda on political issues (McLeod, Kosicki & McLeod, 1994; McCombs, 1994; Walker and Kiwanuka-Tondo, 2003). These areas offer promising research opportunities in the context of AIDS/HIV nationally and globally and can help define public policy discourse.

ACKNOWLEDGMENT

The author would like to thank SAS Institute for its support of this research.

REFERENCES

Aday, L.A. (1994). Health status of vulnerable populations. *Annual Review of Public Health, 15,* 487–509.

AIDS at 25 Special Report. (2006). *Newsweek* (May 15).

Centers for Disease Control and Prevention (CDC). (2005). HIV/AIDS update. A glance at the HIV epidemic. Retrieved June 1, 2006, from http://www.cdc.gov/nchstp/od/news/At-a-Glance.pdf

Data Bulletin #25. (2003). Tracking healthcare costs.

DeSanctis, G., & Poole, M. S. (1994). Capturing the complexity in advanced technology use: Adaptive structuration theory. *Organization Science*, *5*(2), 121-147.

The Datawarehouse Institute Web site. Retrieved July 12, 2006, from http://www.tdwi.org/

Fonkych, K., & Taylor, R. (2005). *The state and pattern of health information technology*. Santa Monica, CA: RAND.

Isaacs, S.L. (2004). Class—The ignored determinant of the nation's health. *New England Journal of Medicine, 351*(11), 137–42.

Kim, W. (2006). Does class trump race? *DiversityInc*., 24–56.

Kovak, C. (2005). Toward a digital health-care ecosystem. *Wall Street Journal*.

LaVeist, T.A. (2005). *Minority populations and health*. San Francisco, CA: Jossey-Bass.

McCombs, M.E. (1994). News influence on our pictures of the world. In J. Bryant, & D. Zillman (Eds.), *Media effects: Advances in theory and research* (pp. 1–16). Hillsdale, NJ: Lawrence Erlbaum.

McLeod, J.M., Kosick, M.G., & McLeod, M.D. (1994). The expanding boundaries of political communication effects. In J. Bryant, & D. Zillman (Eds.), *Media effects: Advances in theory and research* (pp. 123–162). Hillsdale, NJ: Lawrence Erlbaum.

Pitts, M.K., & Jackson, H. (1993). The press coverage of AIDS in Zimbabwe: A five-year review. *AIDS Care, 5*(2), 223–230.

SAS Corporate Web site. Retrieved July 12, 2006, from www.sas.com

Shi, L., & Stevens, G.D. (2005). *Vulnerable populations in the United States*. San Francisco, CA: Jossey-Bass.

Stevens, G.D., &d Shi, L. (2003). Racial and ethnic disparities in the quality of primary care for children: A review of the literature. *Medical Care Research Review, 60*(1), 3–30.

UNAIDS. (2004). *UNAIDS 2004 report on the global HIV/AIDS epidemic*. Geneva, Switzerland: UNAIDS.

Walker, A., Rios, D., & Kiwanuka-Tondo, J. (2003). The band still plays on: A content analysis of HIV/AIDS, African Americans, and Latinos in New England newspapers. In D.I. Rios, & A.N. Mohamed (Eds.), *Brown and black communication: Latino and African American conflict and convergence in mass media* (pp. 93–104). Westport, CT: Praeger.

KEY TERMS

Business Analytics/Intelligence: Use of data and information to drive business actions, including concepts of performance management, definition and delivery of business metrics, data visualization by using online analytical processing (OLAP), dashboards, scorecards, analytic applications, and data mining.

Capitation: Typically associated with managed care; a fixed payment or reimbursement plan that physicians receive based on care delivery to patients.

Cultural Competency: The ability of an organization, group, or individual to work crossculturally in an effort to administer health care services and treatment; this is critical to the patient-physician relationship along with access to appropriate care and quality outcomes.

Diagnostic Related Groups (DRGs): A classification of hospital case types into groups expected to have similar hospital resource use. The groupings are based on diagnoses (*using* ICD), procedures, age, sex, and other attributes.

Health Disparities: Differences in the incidence, prevalence, mortality, burden of disease, or other adverse health conditions that exist among specific population groups.

Healthy People 2010: Objectives for the United States to achieve specified health care status, based on a 10-year timeframe. These objectives include extensive discussion on vulnerable populations, community groups, providers, and so forth, in an effort to improve overall health.

ICD-9 Codes: International Classification of Diseases–9th Division

Managed Care: An arrangement of specific health providers that are contracted through a structured plan for managing service delivery and cost of care.

Socioeconomic Status (SES): A measure of social and economic position that is a function of occupation, income level, and education.

Vulnerability: Propensity of poor health and/or lack of access to health care services.

Application of Wireless Data Grids for Health Informatics

Omer Mahmood
Charles Darwin University, Australia

INTRODUCTION

Distance and inaccessibility create special challenges for health practitioners in rural and remote areas. Health professionals in rural and remote areas face problems such as prescription concerns due to lack of information. This occurs due to a gap in knowledge regarding medications, as health practitioners do not have access to medical history of their patients. Frequent migration of patients in rural and remote communities results in the loss of patients' medical records. In addition, doctors have limited access to therapeutic information, as searching the Internet from most remote communities is expensive and slow. In addition, frequent migration of patients in remote communities results in the loss of patients' histories. This chapter proposes a refined conceptual health information management model based on the model presented by Mahmood (2006). The discussed model is based on the use of Data Grid technology and Data Recharging techniques employed in conjunction with wireless communication technologies to overcome the problems and challenges faced by health practitioners. The model categorizes the health establishments in remote areas into two categories on the basis of geographical characteristics and data access requirements. The discussed model aims to meet information and communication technology (ICT) requirements of health practitioners operating in each recognized category.

INFORMATION SOCIETY, E-HEALTH, TELEMEDICINE, AND HEALTH INFORMATICS

An information society is one in which the main product or a prerequisite to other products is information. In essence, in an information society, the workers' outputs and organizations' successes depend on their ability to utilize information (Knight, 2004). The main recognized sections of information society are information economy, electronic commerce, and e-health (Walsh, 2001).

Mitchell defined e-health as "a new term needed to describe the combined use of electronic communication and information technology in the health sector—the use in the health sector of digital data-transmitted, stored and retrieved electronically—for clinical, educational and administrative purposes, both at the local site and at a distance" (Mitchell, 1999). E-health includes the use of ICT to enable or improve health care service in both clinical and nonclinical sectors. It is composed of telemedicine and health informatics. Telemedicine is direct or indirect delivery of health care or health information. The direct means could include videoconferencing, and the indirect means could include Web sites. In telemedicine, physicians examine patients at a distance by using ICT (Preston, 1993). In relation to telemedicine, health informatics or medicine informatics deals with the resources and devices used to acquire, store, move, and retrieve data to support health care (Wang, Turner & Scott, 2003). So in essence, the application of health informatics is telehealth. This relation is presented in Figure 2.

CATEGORIES OF REMOTE HEALTH ESTABLISHMENTS

The ICT requirements and conditions for health informatics are very different in urban, rural, and remote areas. To set up effective health informatics, it is required to identify and categorize the rural and remote health establishments. The division should be based on the geographical location, mobility requirements of health service providers, current state of ICT infrastructure, and the migration habits of the patients in the recognized establishments. In this case, rural and remote area health establishments are divided into two categories: fixed and mobile remote establishments (Mahmood, 2006). Characteristics and ICT requirements of each establishment are outlined next.

Figure 1. Components of information society (Walsh, 2001)

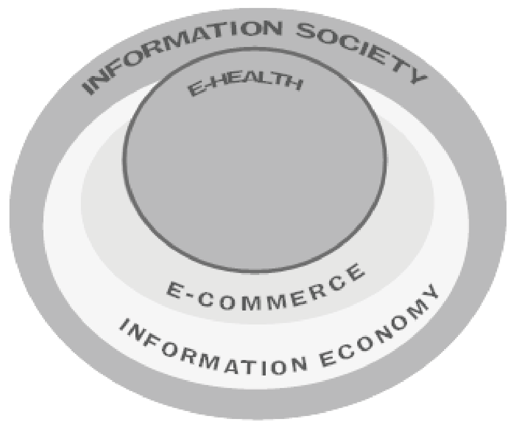
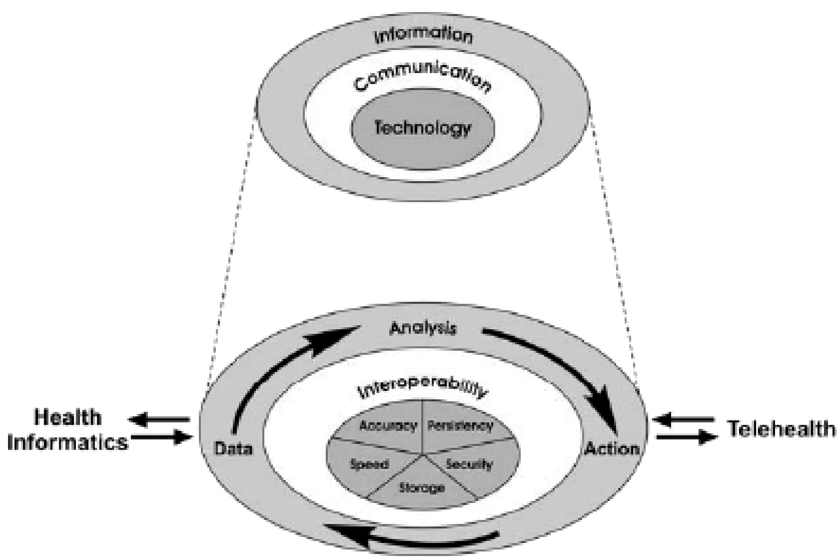


Figure 2. The e-health ICT impact model (Wang et al., 2003)



Fixed Establishments

Fixed remote/rural health care establishments consist of permanent health clinics that cater to the specific but limited health-related needs of local communities. Such establishments require access to limited data set, as the treatments and health services they provide are limited. These establishments do not require frequent synchronization of patient records between rural and urban health establishments due to lower migration of local community members. Semisupportive preestablished ICT infrastructure, such as phone lines, also exists at such establishments.

Portable Establishments

Portable health establishments generally consist of doctors, health practitioners, and district medical officers (DMOs) who visit relatively small communities in remote/rural areas. In such establishments, health practitioners take mobile electronic devices such as a laptop or a pocket PC to access and record the prescription information when undertaking consultations. Practitioners require highly filtered information as they offer limited health services. They have well-defined and limited patient base because of the community size. The data synchronization requirements of such establishments are quite high in relation to fixed establishments due to the high mobility of staff and migration habits of local community members.

The following section discusses the conceptual model. The discussed model targets to meet the requirements of both categories of remote establishments defined previously. It is divided into data grids, data recharging, and wireless Internet service modules (Mahmood, 2006).

HEALTH INFORMATION MANAGEMENT SYSTEM

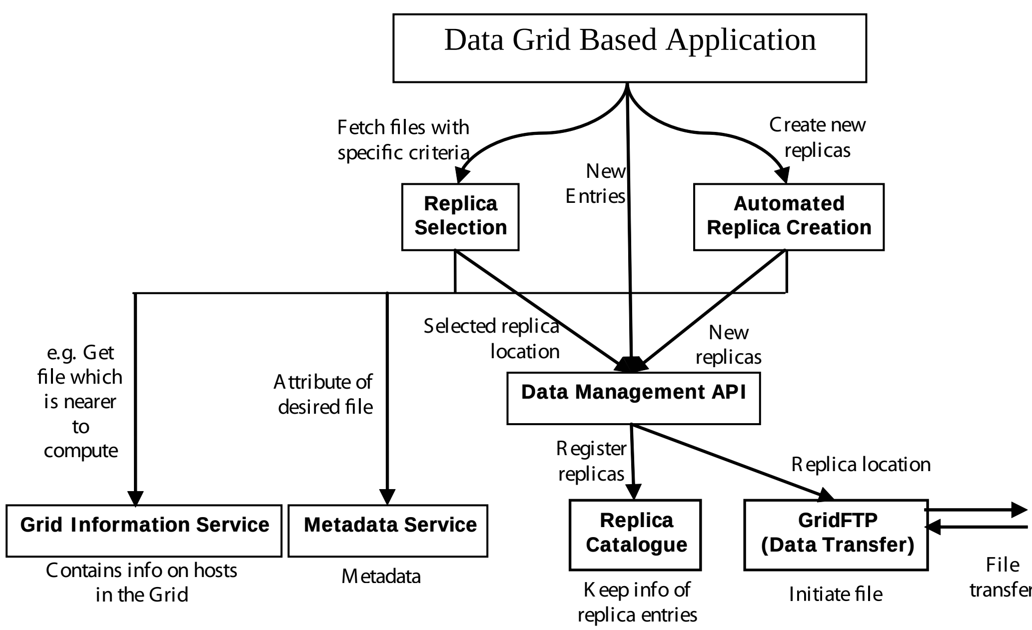
Health information management systems for health informatics need to maintain patients' health information and medical records of diagnosis, treatments, and procedures. Data grids have been proposed to meet data storage, movement, retrieval, and replication requirements. Data grids provide flexible infrastructure to share Internet-connected resources including computers, data storages, software applications, sensors, and instrument resources and wireless equipment (Tian, 2003). It is proposed to use data recharging techniques based on user profiles for information filtering, customisation, and prefetching. The application of data prefetching and filtering assists to reduce the record management and migration overhead (Mahmood, 2005). For cost-effective, robust, and easily maintainable networking, a combination of commercially available wireless LAN, GPRS (GSMWorld, 2006), and satellite-based connections is suggested.

Data Grids

Data grids enable the exchange of data from different sources. Existing in physically diverse locations, they hide all the differences from the users by enabling them to access data as if the data are on the local disk. Due to inbuilt features such as replication management and security, data grids offer a number of advantages over the traditional data access methods. These features of data grids make them most suitable for data exchange and sharing. Such features play a pivotal role if the data are distributed over a wide geographical area, as in case of health service providers operating in remote/rural communities where the patient's records exist at varied diverse location.

Drug researchers are currently using data grids to gain access to comprehensive patient records that meet designated characteristics such as age, race, geographical areas, and so forth. Such data could come from large disparate databases located in hospitals at many geographic places (Tian, 2003; Kermani, 2003). The

Figure 3. GridFTP and data selection scenario (Kim, 2002)



main internal services of data grids include GridFTP, a high-performance, secure, robust data transfer mechanism; Replica Catalogue, a mechanism for maintaining a catalogue of dataset replicas; and Replica Management, a mechanism to bind Replica Catalogue and GridFTP components. Replica Management service enables grid applications to create and manage replicas of datasets. Modules of replica management include creating new copies of complete or partial collection of files, registering these new copies to replica catalogue, and enabling users and applications to query and search the catalogues to locate all existing copies.

Data Recharging

Data recharging or data prefetching enables any device to plug into the wired or wireless network at any location for any amount of time and, as a result, end up with more useful data than it had before. The data recharging service constitute of a network of profile managers is located throughout the system network. Profile managers are responsible for retrieving required data from data sources for delivery to specific devices based on their recharging profiles (Cherniack, Franklin & Zdonik, 2001). In the discussed system, the profile managers interact directly with the underlying data grid system, requesting and, if necessary, reformatting or transforming the data (e.g., to XML, WML, or XHTML formats).

The system utilizes two types of profile managers: user profile and device profile managers. The user profile manager is used to filter the information in accordance with the expertise and requirements of the user; in this case, the health practitioner (Mahmood & Zomaya, 2006). The user profile manager builds and maintains the user's profile on the basis of job description of the user, offered services description by the health establishment from where the user operates, and the filters deployed by the user. This prefiltering technique is used to save time and to better utilize the limited network resources. The user profiles evolve over time and are somewhat static. Once the user profile manager filters the data, it is sent to the device profile manager. The device profile manager uses user mobile profile (UMP) to further filter the contents on the basis of capabilities of the mobile device the user is currently using. These two profiles are used to prefetch the information on a user's mobile device so the user can operate even in the absence of network coverage.

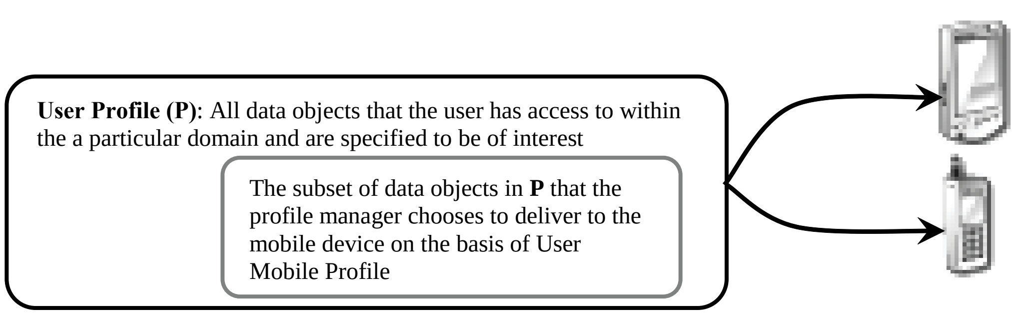
Striped Data Transfer (SDT) module of data grids can automatically strip defined data types resulting in smaller file sizes. SDT can also better utilize the available bandwidth by applying parallel transfers (Allcock et al., 2001). Moreover, STD, with the replica catalogue and replica management components, can also move the data closer to a user's geographical location. This feature significantly reduces bandwidth requirements, as the system can be configured to move the data when the network load is less, resulting in better use of network resources.

Hardware Based System Components

Within an ICT domain, the main characteristic of urban health establishments is the existing availability of stable telecommunication infrastructure supporting the availability of high network bandwidth. However, a new challenge arises if it is required to exchange data between urban and remote/rural fixed or between rural/remote fixed and portable establishments, since preestablished ICT infrastructure usually lacks in such geographical areas.

To meet the bandwidth requirements and share information with both types of health establishments,

Figure 4. Use of profiles for data recharging (Cherniack et al., 2001)



two networking solutions are discussed that implement and make use of a combination of different wireless communication technologies. For information exchange between remote/rural fixed and portable establishments, wireless local area network (WLAN), GPRS, or satellite connection could be used. The unique blend of communication technologies for each category of establishment is tailored to keep the setup, running, and maintenance costs low while providing dynamic, robust, secure, scalable, and customized architecture.

Fixed establishments. The choice of technology primarily depends on the aerial distance between the establishments. If the distance between fixed establishments is less than 23 kilometers, then it is proposed to use wireless bridges to extend the range of 802.11g WLAN network. Wireless bridges can connect locations up to 23 km apart, and can carry voice and data traffic at speeds up to 200Mbps. Wireless bridges use Radio Frequency (RF) wireless technology that can be secured by using encryption at each end. Table 1 outlines the data transfer speed and area coverage for each IEEE 802.11 WLAN standard when used in conjunction with range extenders (Long Range Team, 2006).

Portable establishments. In geographical areas covered by cellular networks such as Global System for Mobiles (GSM), EDGE, UMTS, TDMA, or Code Division Multiple Access (CDMA), data-added services could also be used. CDMA supports high-quality voice clarity and a bandwidth up to 14.4Kbps for data transfer, GPRS support up to 171.2Kbps of data transfer rate, while 3G supports up to 2Mbps for stationary systems. Like CDMA, GPRS and 3G enable users to use FTP service to transfer files, browse the Web, send and receive e-mails, and synchronize information with the server. It is recommended to use cellular data-added services only if the distance between the fixed and portable remote establishment is more than 23km, as the running cost would be much higher and the available bandwidth is significantly lower than WLAN.

However, in situations where the distance between urban, rural/remote fixed and portable establishments is greater then 23km and the establishment is not serviced by a cellular network, then it is recommended to use satellite connection. Mobile Packet Data Service (MPDS) supports digital voice, fax, and data exchange anywhere, anytime, and securely by using satellites. Such services provide up to 128Kbps inbound and up to 2Mbps outbound data transfer rate. The data transferred over a satellite network incur high running costs. In case of a doctor visiting a remote community, a Point-to-Point connection can be established where satellite is used to interconnect the fixed establishment with portable establishment. Use of satellites to transfer data is the most expensive method; therefore, this should be engaged only if no other networking method can be implemented.

CONCLUSION

This chapter proposes a conceptual model to meet data management and communication challenges faced by health practitioners operating in rural and remote communities due to lack of ICT infrastructure and frequent movements of patients. A blend of wireless communication technologies along with data grid applications and data recharging techniques are proposed to meet the health practitioners' requirements. In the model, the rural and remote health establishments are divided into two categories on the basis of their geographical characteristics and data access requirements. The model recommends using different wireless networking technologies on the basis of geographical characteristics of health establishments. It is proposed

Table 1. Comparison of 802.11 standards with range extenders

IEEE	Speed	Indoor Range	Outdoor Range*
802.11b	11 Mbps	100 meters	up to 23 km
802.11a	54 Mbps	100 meters	up to 5 km
802.11g	54 Mbps	100 meters	up to 23 km
802.11n (Draft)	200+ Mbps	150 meters	up to 23 km

*Note: * By using amplifiers, range extenders, bridges, and repeaters* to use Data Grids to create data backups, replications, and movements. Moreover, because of the built-in replica management feature of Data Grids, the patient's records can be replicated and moved automatically in response to his or her physical movements. The discussed model relies on the use of data recharging and data filtering techniques to cache and prefect the most relevant information to the user's device, thus enabling the user to operate efficiently even in the absence of network availability.

REFERENCES

Allcock, B., et al. (2001). *Data management and transfer in high-performance computational grid environments.* Chicago: University of Chicago, Department of Computer Science & the Computation Institute. Retrieved May 7, 2006, from http://www.globus.org/alliance/publications/papers/dataMgmt.pdf

Cherniack, M., Franklin, F., & Zdonik, S. (2001). *Expressing user profiles for data recharging*. IEEE Personal Communications: Special Issue on Pervasive Computing. Retrieved from http://citeseer.ist.psu.edu/cherniack01expressing.html

GSMWorld. (2006). *GRPS platform*. GSM World Web site. Retrieved November 8, 2006, from http://www.gsmworld.com/technology/gprs/index.shtml

Kermani, M. (2003). *Practice and experience in clinical deployments of storage grids for the transmission and storage of diagnostic imagery*. Proceedings of the GlobusLife Sciences Workshop, San Francisco, California.

Kim, E. (2002). *Data management in Globus*. KISTI Supercomputing Center. Retrieved May 7, 2006, from http://testbed.gridcenter.or.kr/kor/seminar/seminar2/eskim74.ppt.

Knight, A. (2004). *Internet publishing and digital rights: The changing balance between access and ownership.* CSA Web site. Retrieved October 18, 2006, from http://www.csa.com/discoveryguides/internet/gloss_f.php

Long Range Team. (2006). *Long range wireless solutions*. Simply Wireless Australia Web site. Retrieved May 7, 2006, from http://simplywireless.com.au/longrange.htm

Mahmood, O. (2005). *Adaptive profile driven data caching and prefetching in mobile environment*. Sydney, Australia: University of Sydney, School of Information Technologies. Retrieved from http://hdl.handle.net/2123/714

Mahmood, O. (2006). Blend of technologies to facilitate health practitioners in remote and rural Australia. *Journal on Information Technology in Healthcare, 4*(6), 366–372.

Mahmood, O., & Zomaya, A. (2006). Adaptive profile driven data caching and prefetching in mobile environment. *Journal of Ubiquitous Computing and Intelligence, 1*(1), 1555–1334.

Mitchell, J. (1999). *From telehealth to e-health: The unstoppable rise of e-health*. John Mitchell & Associates for the Federal Australian Department of Communications, Information Technology and the Arts (DOCITA).

Preston, J. (1993). *The telemedicine handbook. Improving health care with interactive video*. Danvers, MA: PictureTel Corporation.

Tian, H. (2003). *Grid computing as an integrating force in virtual enterprises.* Massachusetts Institute of Technology. Retrieved May 7, 2006, from http://wirelessgrids.net/docs/hongfei.pdf

Walsh, M. (2001). *National telehealth plan for Australia and New Zealand*. Health Online.

Wang, X., Turner, L., & Scott, R.E. (2003). Technology solutions—Saving the day? In *Proceedings of the 6th Annual Can Soc of Telehealth Conference*, Halifax. Nova Scotia.

KEY TERMS

E-Health: A collective use of electronic communication and information technology in the health sector to electronically transmit, store, and retrieve digital data, both at the local site and at a distance, for clinical, educational, and administrative purposes. It is composed of telemedicine and health informatics.

Health Informatics: The use of devices and resources to collect, store, move, and retrieve data to support health care. One application of Health Informatics is telehealth.

Health Information: Recorded information in any form or medium about the health of a particular individual. It is created or received by a health practitioner or provider. It consists of past and present health condition of as well as past, present, and future treatments conducted on an individual.

Health Information Management System: An information system consisting of computer hardware and software, procedures, and processes that are specifically designed and implemented to store, maintain, collect, process, represent, and manage information specific to the health care domain.

Information Society: One in which the main product or a prerequisite to other products is information. In essence, in information society, the workers' outputs and organizations' successes depend on their ability to utilize information.

Telehealth: The distribution of health-related information or services by using telecommunication methods such as health-related advice over the phone; distribution of health-related information by using the Web; use of videoconferencing, online chat, online discussion boards, and exchange of medical records over the Internet or by using fax.

Telemedicine: Direct or indirect electronic delivery of health care or health information.

Applications of Data Mining in the Healthcare Industry

John Wang
Montclair State University, USA

Xiaohua Hu
Drexel University, USA

Dan Zhu
Iowa State University, USA

INTRODUCTION

The workflow of health care organizations involves the generation and collection of various kinds of data relating to clinical practices, clinical trials, patient information, resource administration, policies, and research. Traditionally, statistical techniques are used to derive some operational information from the data. Data mining, a new method, provides the opportunity to derive, in an exploratory and interactive manner, valuable health care knowledge in terms of associations, sequential patterns, classifications, predictions, and symbolic rules. Such inductively-derived health care knowledge can provide strategic insights into the practice delivery of health care.

As the medical field expands, it is the duty of each physician to evaluate and protect each patient from diseases, side effects, and medical mishaps. Armed with a scalpel, stethoscope, and other accruements, physicians are now armed with data mining as a tool for expanding their knowledge base. Data mining is available to every aspect within the health care industry. It is multifaceted and used in areas like insurance to detect fraud, the pharmaceutical industry to evaluate side effects of drugs, and even detection of certain diseases based on genetics.

BACKGROUND

The progress of data mining has been around for decades. Although data mining is an evolution of a field that can be traced back many years, the terminology itself has been only been introduced in the 1990s. By using pattern recognition technologies and statistical and mathematical techniques to sift through warehoused information, data mining helps analysts recognize significant facts, relationships, trend, patterns, exceptions, and anomalies. The use of data mining can advance a company's position by creating a sustainable competitive advantage. Data mining is the science of managing and analyzing large datasets and discovering novel patterns (Davenport & Harris, 2007; Olafsson, 2006; Wang, 2006).

Data mining is becoming more prevalent in the health care industry because of the vast quantities of data stored in a multitude of medical systems, more specifically systems of health care providers, hospitals, and other medical institutions (Veletos, 2003). Recognizing patterns of data in order to discover valuable information, new facts, and relationships among variables are important in making business decisions that would best minimize costs, maximize returns, and create operating efficiency without compromising the quality of patient care.

The growth in data mining and its advantages over the traditional methods of collecting data in the industry, both for reducing costs and providing patient care, will be discussed. The bulk of the material focuses on various applications of data mining in the health care industry under three major areas. These areas include: data mining in health insurance plans; data mining in clinical care, and data mining in health care administration. Also, emerging future trends in data mining in the health care industry will also be explored.

MAIN FOCUS

Data Mining in Health Insurance Plans

Financially oriented applications, such as the actuarial cost-risk models utilized by health insurance companies, as well as models used to track and identify potential claim fraud, have been the primary focus of data mining in the health care industry. Along with these applications, predictive models have also been applied to issues that impact cost, such as predictions of length of stay, forecasting treatment costs, total charges, and even mortality. These applications contribute significantly to the driving force that has helped this industry evolve to the gigantic sector that it is in the market: how can health care organizations minimize cost, raise its quality of service, and maintain its competitive edge at the same time.

Fraud Detection

Research indicates that health care fraud and abuse costs the U.S. as much as $100 billion per year. That equates to about 10% of annual spending on health care. Numbers like that represent a huge red flag for action in this industry. Fraudulent and abusive behavior has become a serious problem. It is an area that requires significant attention and insight in order to prevent the unnecessary expenses associated with fraudulent transactions. Fraud detection takes on many forms including: claim fraud, premium avoidance fraud, and indemnity fraud (as in the case of disability or workers compensation insurance).

Claim fraud is currently the most predominant form of fraudulent activity and is a major concern of health care organizations. Medical claim fraud takes place when medical providers deliver unnecessary treatments and services or file claims to the insurance companies for services that have not been rendered in actuality. Due to the bulk of medical claims being processed side by side with a great multitude of insured patients and medical providers, detection of fraudulent claims poses a great challenge to health insurance companies. There is a considerable degree of difficulty involved in discerning whether a particular transaction is valid or illegitimate due to the enormous quantity of claims being processed.

Another type of medical fraud mentioned above is *premium avoidance fraud* which can occur under a variety of circumstances. Some examples include: (1) A policy holder attempting to reduce premiums by providing misinformation to the insurance carrier, such as fewer employees or lower payroll; (2) A policy holder submits incorrect classification to qualify for lower premium rates; (3) A policy holder changes insurance to cover injured employee and falsifies accident date; or (4) An employee changes a date-of-birth for a dependent that would otherwise not qualify for coverage.

Indemnity fraud, on the other hand, occurs when an insured customer receives workers compensation indemnity benefits (exemption from paying incurred liabilities) when in reality, they are no longer qualified to do so due to reasons such as capability of returning to work or discontinuation of medical treatments.

All of the forms of fraudulent activities mentioned above represent a consequential opportunity for high-value data mining applications in the insurance industry.

Yang and Hwang (2006) proposed a process-mining framework that uses clinical pathways to distinguish fraudulent and abusive cases from normal ones. The framework for their research consisted of first breaking the health care industry down by service groups. Then, for each care service, a set of features is identified and an inductive model is developed to detect suspicious claims for each particular service group. All of these methods reduce the workload of human experts.

Policy Holder Retention

In order to sustain desired revenues, programs desired to retain current members are vital to a medical insurance organizations. A significant threat to the insurance industry is a reduction in the number of policy holders. It has been recognized that retention of existing customers is about 5-10 times more cost effective than acquiring new ones (Haux, Elske, Werner, & Knaup, 2002). Therefore, it is considered wise for insurance companies to distinguish members who are at risk of switching their membership to another provider.

Data mining can be used to identify the potential loss of clients switching to other companies. Historical data contains usage patterns and other important customer characteristics that when discovered can be used to identify satisfied or unhappy policyholders. These correlations between historical data values are very complex. Predicting policyholders who are at risk of changing their service to another provider with a high level of accuracy is extremely complicated, and it requires a lot of experience. These, nonetheless, are very powerful profiling techniques.

New Policy Holders

Data mining can also be used to help identify potential new customers. By using historical data to predict future behavior, the marketing departments of insurance companies can determine potential policyholders whose behavior will be less risky then others. These high value, low risk potential policyholders can be approached with special packages developed for them in the hopes of attracting these customers from other companies.

Liability Profiling

Being able to predict the payout amounts that insurance companies would have to make is extremely important for both the long term financial health of the company, as well as being able to pay the company's current obligations. By predicting policyholders' behavior and risk, insurance companies can use these data mining models to construct the overall risk of a policyholder. Once the policyholder has a risk rating, the appropriate premium can be charged and the fee can be collected. Also, if the insurance company determines that the company, as a whole, will be paying out more than originally estimated, the company can adjust rates and assets to avoid any capital short fall.

Customer Service

Data mining is also used to predict when the call center will experience an increase or a decrease in volume. This allows for better staffing and better allocation of resources. Data mining is also used for the planning of resources when certain scenarios occur. For example, after an earthquake or wild fire, the insurance company can increases the call center staff to handle the increase in the volume caused by the disaster (Silver, Sakata, Su, Herman, Dolins, & O'Shea, 2001).

Cross Product Marketing

Once an insurance company has a profile on a customer, they often use the information to market other products that the insurance company provides, to the same customer. For example, once an insurance company provides health insurance for a customer, the company will often try to sell life insurance to the same customer. Data mining allows for greater predictability of the client's future needs.

Data Mining in Clinical Care

Data mining provides the ability to correlate differences in treatment with significant impact on the clinical outcomes. These tools can automatically identify the subgroup within a population, and provide descriptive statistics regarding how each subgroup compares with the population as a whole or with other subgroups.

Epidemiology

Interactive data mining, also known as *knowledge discovery* tools, can be used to identify critical variation and define priorities (Friedman & Fisher, 1999). Data mining in clinical care also uses pattern recognition to identify significant variations. This process allows doctors to identify the types of drugs that are associated with adverse effects. By combining data mining with statistics such as hypothesis testing, clinical researchers are able to improve clinical outcomes. This is achieved by comparing the sample result to the population, which can thereby determine the overall effectiveness of the drugs.

Patient Profiling

Using data mining techniques such as pattern recognition, physicians and patients can learn useful clinical trials of new treatments. Pattern recognition can also be used to prevent diseases before they occur. For example, a physician can identify patients who are at a great risk for having a heart attack before they even show any signs of the disease. This can be determined just by the pattern of the lifestyle. This test can be performed, and if there is an issue found, then heart disease medication could be prescribed to prevent a heart attack (Kathy, 1998).

The most popular data mining techniques used by physicians and pharmaceutical companies are *neutrality* tests, also known as *market based* tests. Market based tests utilize information from numerous local regions to construct a null distribution based on the variability characteristics of certain samples (Vasem & Primmer, 2005). This allows many variables to be considered especially when the size of the sample is small, large, or evaluated over time against the use of other drugs.

When long term side effects are considered, many faults or issues may arise. Since 1999, many vaccines, which are preventative and necessary in children, were pulled from the market. As vaccination increases paralleled increases in population many side effects would be noticed. Iskander, Pool, Zhou, and English-Bullard (2006) discovered that during 1991 and 2003, vaccinations tripled along with a proportionate number of adverse effects. With the use of market based data mining techniques, physicians are able to evaluate syndromes, symptom interactions, and intervaccine interactions, and evaluate potential adverse conditions with vaccinations. Medical data divided items into two categories: procedure and diagnosis. Unlike market base data, the association between procedure and diagnosis reveal cause and effect relationships (Doddi, Marathe, Ravi, & Torney, 2001).

Disease Management

Data mining applications are used to analyze treatments for particular diseases in order to visualize its possibility for further clinical enhancement. One scenario where data mining techniques could be applied would be a study of the relationships between two ailments and finding the most effective treatment. For example, a given health care provider could set up a regional test to examine all occurrences of asthma and allergies, find the connections between these two diseases, and look for the most efficient and effective way to combat them. The data which would be collected, studied, and compared could include types of over-the-counter (OTC) drugs or prescription medication such as antihistamines, as well as more holistic treatments such as taking a spoonful of honey daily while carefully taking into consideration the placebo effect. In line with Borok (1997), monitoring environmental factors such as pollution, pollen count, and airborne microorganisms, to name a few, would also be crucial to the study since they are also causes of the specified ailments. Through a thorough examination of these factors and their relationships to the different treatment regimes (pattern recognition), data mining could render a best treatment alternative.

Data mining could also be used in preventing the spread of a contagious ailment and forestalling the diffusion of an impending epidemic. According to the Centers for Disease Control and Prevention, the recurrence and resurgence of a disease which otherwise would have been completely eradicated stems from the tendency of patients to discontinue medication (even though the dosage has not been completed) the moment symptoms of the illness cease to exist (Srivastava & Singh, 2004). Through a rigorous analysis and careful monitoring, data collected from the study can be used to graphically represent the spread and progress of the disease. Hence, researchers would be able to discern patterns of the illness and take the necessary actions to address them.

Milley (2000) pointed out that along with the prevention of diseases, data mining can also be used to minimize the effect of many diseases before they progress further. For example, if there is a strong correlation between people with high blood sugar and blindness, doctors could require more frequent exams of those patients with high blood sugar level and check for blindness, than for those other patients who do not have high blood sugar level. This predictive model will keep doctors ahead of the disease and improve the overall quality of life fore most patients.

Data Mining in Health Care Administration

Data mining models are often used to support health care administration functions by predicting the following:

- Staff recruiting and staff retention
- Length of time an operating room will be required
- Number of beds that will available
- Expected return on investment per procedure
- Amounts and types of medicines on hand
- Number of ambulances
- Mortality rate per procedure and by doctor

Hospitals use this information for allocating resources and making decisions. As in most data mining models, validating data, testing algorithms, and clearly understanding methods make for accurate models. In addition to the standard method of data mining tools, according to Biafore (Biafore, 2003), hospitals often use utilization predictors. These tools allow hospitals the ability of knowing ahead of time which health care resources will be needed (both from the point of view of an entire member population and each individual member). These predictive solutions offer the capability of combining disparate data sources including both financial and clinical data to achieve accurate forecasts of resource utilization, across a wide variety of resource types.

Besides being able to utilize their facility optimally, hospitals and clinical organizations are often plagued by unpaid bills, collection agency fees and outstanding medical testing costs. Scheese (1998) estimated that, as a matter of fact, hospitals typically end up paying 30% to 50% of recovered bad-debt revenue to outside collection agencies.

FUTURE TRENDS

More Specific Package

Nowadays, many data mining packages available on the market are mainly for business use and are either too expensive or not specific enough to handle medical information (Liao & Lee, 2002). Characterized by a multitude of records and numerous attributes, health databases have become a challenge for data miners in their pursuit of developing applications and utilizing methodologies which would cater specifically to the health care industry. Applying more specific data mining to the entire health care industry would be the best way for physicians to evaluate the probability of patient interactions or reactions to certain treatments or diagnosis.

Appropriate Information Infrastructure and End-user Tools

The bottleneck for effective application of data mining in health care industry is lack of appropriate information infrastructure and end-user tools (Augen, 2003; Olafsson, Li, & Wu, 2006). Building infrastructures to support the free exchange of valuable health care data is vital for health researchers. By creating such infrastructures, there will be more data available to more researchers who can use data mining techniques to nugget out valuable hidden knowledge. With the proper infrastructure in place, doctors can use historical data and learning technology to find the best treatments for their patients. We believe in the near future, data mining software will become a major component of tomorrow's medical kit and a leading ally in the fight against illness and disease.

CONCLUSION

Data mining has definitely made an impact in the health care profession. Through the many data mining techniques discovered, medical experts have been able to improve the quality of care and service they provide to their patients. Data mining has become so popular that it can be used on the business side of health care to help reduce costs as well as the theoretical side in finding hidden knowledge in medical records. Experts have identified future trends and what needs to be done in order for them to be realized.

With data mining, hospitals would be able to monitor the data in a more sophisticated and intelligent way so that they can look at events that are occurring and find things they might not have thought of before. This new way of tracking data is more advanced than traditional analysis because they can find patterns that indicate problems to come. Hospitals have the chance to stay one step ahead of problems. Moreover, if they can be proactive, hospitals can avoid bigger problems. After all, timely information and early intervention is the key for preventing diseases. Considering the vast progress data mining is already responsible for in the medical field, one can only be excited to wonder what other great benefits will come about because of it.

REFERENCES

Augen, J. (2003). The increasing importance of large-scale information systems for health care R&D. *Drug Discovery Today*, *8*(8), 329-331.

Biafore, S. (2003). Predictive solutions bring more power to decision makers. *Health Management Technology*, *20*(10), 12-15.

Borok, L. (1997). Data mining: Sophisticated forms of managed care modeling through artificial intelligence. *Journal of Health Care Finance*, *23*(3), 20-36.

Davenport, T., & Harris, J. G. (2007). *Competing on analytics: The new science of winning*. Harvard Business School Press.

Doddi, S., Marathe, A., Ravi, S.S., & Torney, D.C., (2001). Discovery of association rules in medical data. *Medical Informatics & the Internet in Medicine*, *26*(1), 25-33.

Friedman, J., & Fisher, N. (1999). Bump hunting in high-dimensional data. *Statistics and Computing, 9*(2), 123-143.

Haux, R., Elske, A., Werner, H., & Knaup, P. (2002). Health care in the information society: A prognosis for the year 2013. *International Journal of Medical Informatics*, *66,* 3-21.

Iskander, J., Pool, V., Zhou, W., & English-Bullard, R. (2006). Data mining in the U.S. using the vaccine adverse event reporting system. *Drug Safety*, *29*(5), 375-384.

Kathy, K. (1998). Data mining: Digging for health care gold. *Insurance & Technology, 23*(2), 4.

Liao, S., & Lee I. (2002). Appropriate medical data categorization for data mining classification techniques. *Medical Informatics & the Internet in Medicine, 27*(1), 59-67.

Milley, A. (2000). Health care and data mining. *Health Management Technology, 21*(8), 44-46.

Olafsson, S. (2006). Introduction to operations research and data mining. *Computers & Operations Research, 33*(11), 3067-3069.

Ólafsson, S., Li, X., & Wu, S. (2008). Operations research and data mining. *European Journal on Operational Research*, *187*(3), 1429-1448.

Scheese, R. (1998). Data warehousing as a health care business solution. *Health care Financial Management*, 56-59.

Silver, M., Sakata, T., Su, H. C., Herman, C., Dolins, S. B., & O'Shea, M. J. (2001). Case study: How to apply data mining techniques in a health care data warehouse. *Journal of Health care Information Management, 15*(2), 155-164.

Srivastava, A., & Singh, V. (2004). *Deriving interpretable rules for financial outliers in rush hospital data* (Unpublished internal technical report). Hitachi America, Ltd.

Vasem, A., & Primmer, C. R. (2005). Challenges for identifying functionally important genetic variation: The promise of combining complementary research strategies. *Molecular Ecology, 14*(12), 3623-3642.

Veletos, A. (2003). Getting to the bottom of hospital finances. *Health Management Technology*, *24*(8), 30.

Wang, J. (2006). *Encyclopedia of data warehousing and mining* (1st ed.). Hershey, PA: Idea Group Reference.

Yang, W., & Hwang, S. (2006). A process-mining framework for the detection of health care fraud and abuse. *Expert Systems with Applications, 31*(1), 56-68.

KEY TERMS

Classification Regression Trees (CART): A mapping of observations about an item to conclusions about the item's end value.

Clinical Pathways: Integrated care pathways are multidisciplinary care plans where diagnosis and treatment are performed by medical teams.

Data Mining: Also known as knowledge discovery in databases (KDD), data mining is the process of automatically searching large volumes of data for patterns. Data mining is a fairly recent and contemporary topic in computing.

Data Mining Surveillance System (DMSS): Uses data mining techniques to discover unsuspected, useful patterns of infections and antimicrobial resistance from the analysis of hospital laboratory data.

Machine Learning: Concerned with the development of algorithms and techniques, which allow computers to "learn."

Neural Networks: Also referred to as artificial intelligence (AI), which utilizes predictive algorithms.

Pattern Recognition: The act of taking in raw data and taking an action based on the category of the data. It is a field within the area of machine learning.

Applying Cross Approximate Entropy to Blood Oxygen Saturation and Heart Rate from Nocturnal Oximetry in Screening for Obstructive Sleep Apnea

Daniel Álvarez
University of Valladolid, Spain

Roberto Hornero
University of Valladolid, Spain

María García
University of Valladolid, Spain

Félix del Campo
Hospital Universitario del Rio Hortega of Valladolid, Spain

Miguel López
University of Valladolid, Spain

Carlos Zamarrón
University of Valladolid, Spain

INTRODUCTION

Obstructive sleep apnea (OSA) syndrome is a major sleep-related breathing disorder, affecting 1% to 5% of adult men in western countries (Young, Peppard, & Gottlieb, 2002). This disease is now the most common respiratory referral to many sleep units (Douglas 2002). OSA syndrome is associated with excessive daytime sleepiness, neurocognitive deficits, and psychological problems. A significant number of patients with suspicion of OSA present mild to severe symptoms of depression (Schwartz, Kohler, & Karatinos, 2005). OSA syndrome is also a risk factor for cardiovascular disease morbidity and mortality (Leung and Bradley 2001). People with OSA have higher risk of sudden death due to cardiac causes during sleep than the general population (Gami, Howard, Olson, & Somers, 2005). Furthermore, subjects with OSA present an increased risk of being involved in traffic and work accidents (Barbé, Pericas, Muñoz, Findley, Anto, & Agustí, 1998).

The recommended diagnostic test for OSA is overnight polysomnography (PSG) (Polysomnography Task Force, 1997), which records neurophysiological and cardiorespiratory signals to analyze sleep and breathing. However, its relative high cost and complexity, along with the large number of people suspected of having this disease, limit its capacity as a diagnostic test for OSA (Whitelaw, Brant, & Flemons, 2005). Due to its simplicity and low cost, *portable monitoring* has been proposed as an alternative method to PSG in the diagnostic assessment of patients with suspected sleep apnea (Flemons & Littner, 2003).

Nocturnal oximetry provides different quantitative indexes commonly used by physicians in screening for OSA. Oxygen desaturation indexes (ODIs), which measure the number of dips in the oxygen saturation signal below a certain threshold, and the cumulative time (CT) spent below a certain saturation level, are the most widely used (Chaudhary, Dasti, Park, Brown, Davis, & Akhtar, 1998; Netzer, Eliasson, Netzer, & Kristo, 2001). However, their diagnostic sensitivity and specificity vary in a wide range between different studies and there is not a consensus on their definition (Netzer et al., 2001). Although oximeters usually provide both blood oxygen saturation (SaO_2) and heart rate (HR), classical indexes like ODIs or CTs are defined exclusively for SaO_2. Few studies have assessed the diagnostic accuracy of HR from nocturnal oximetry to diagnose OSA (Zamarrón, Gude, Barcala, Rodríguez,

& Romero, 2003; Zamarrón, Pichel, & Romero, 2005; Zamarrón, Romero, Gude, Amaro, & Rodríguez, 2001). The study developed by Zamarrón et al. (2003) showed that frequency characteristics from HR signals improved results obtained processing SaO_2 individually. In a later research by Zamarrón et al. (2005), both signals were processed together to quantify the relationship between oscillations in SaO_2 and HR by means of the frequency coherence function. In the present study, we estimate the *asynchrony* between SaO_2 and HR from nocturnal oximetry by means of a nonlinear measure called *cross approximate entropy (cross-ApEn)*. Previous studies have demonstrated the usefulness of nonlinear analysis in the diagnosis of OSA syndrome (Álvarez, Hornero, Abásolo, del Campo, & Zamarrón, 2006; Hornero, Álvarez, Abásolo, del Campo, & Zamarrón, 2007).

BACKGROUND

Similarities between two signals or variables have been traditionally quantified by means of linear methods, like the cross-correlation function or spectral estimates based on cross-spectrum, especially *coherency*. The magnitude squared coherence is a normalized cross-spectral density function that gives a measure of the degree of linear correlation between each of the individual frequency components of two signals (Hampson & Mallen, 2006). It has been widely applied to quantify functional interactions between different brain areas (Nunez et al., 1997). However, these classical estimates are often inadequate to satisfactory identify visually apparent instances of significant association or correspondence between paired series (Liu, Pincus, Keenan, Roelfsema, & Veldhuis, 2005). Statistical estimation methods are nontrivial and their interpretation could be problematic. Moreover, standard spectral estimators can be inconsistent and biased, especially in the presence of outliers and nonstationarities (Pincus, 2000). Particularly, coherency estimates are not suitable to characterize nonstationary signals and only capture linear relations between time series (Porta, Baselli, Lombardi, Montano, Malliani, & Cerutti, 1999; Stam & van Dijk, 2002).

A recently developed measure of asynchrony between time series, the *cross-ApEn*, has demonstrated to be complementary and often superior to both spectral and correlation based methods (Pincus, 2000). *Cross-ApEn* provides substantially more general or robust measures of feature persistence than both correlation and spectral based methods (Pincus, Mulligan, Iranmanesh, Gheorghiu, Godschalk, & Veldhuis, 1996). Cross-ApEn is a family of statistics proposed by Steven M. Pincus, suggesting its application to physiological signals like HR and respiratory rate (Pincus & Singer, 1996). It has been mainly applied to study hormone time series dynamics, like luteinizing and follicle-stimulating hormones (Veldhuis, Iranmanesh, Mulligan, & Pincus, 1999), luteinizing hormone and testosterone (Veldhuis, Pincus, García-Rudaz, Ropelato, Escobar, & Barontini, 2001), or adrenocorticotropic hormone and cortisol (Liu et al., 2005). Other entropy-based measures of asynchrony have been also applied to cardiovascular variability signals, pointing out their advantages over the linear approach (Porta et al., 1999; Porta et al., 2000).

This study was aimed to assess whether a measure of asynchrony between SaO_2 and HR from nocturnal oximetry using *cross-ApEn* could help in the diagnosis of OSA syndrome.

CROSS APPROXIMATE ENTROPY ANALYSIS TO HELP IN THE OSA DIAGNOSIS

Polysomnography and Oximetry Studies

A total of 187 patients with suspicion of OSA syndrome were studied. Sleep studies were carried out in the Sleep Unit of Hospital Clínico Universitario in Santiago de Compostela, Spain. The Review Board on Human Studies approved the protocol, and all subjects gave their informed consent to participate in the study. A nocturnal oximetry study (pulse oximeter Criticare 504, CSI, Waukesha, WI, USA) was performed simultaneously with a conventional PSG (Polygraph Ultrasom Network, Nicolet, Madison, WI, USA). If a subject had less than 3 hours of total sleep, the sleep study was repeated.

The PSG register was analyzed according to Rechtschaffen and Kales rules (Rechtschaffen & Kales, 1968). An *apnea event* was defined as a cessation of airflow for more than 10 seconds, whereas hypopnea was defined as the reduction of respiratory flow for at least 10 seconds accompanied by a 4% or more decrease in the saturation of hemoglobin. An average apnea-hypopnea index (AHI) derived from the PSG study was calculated for hourly periods of sleep. Subjects with an AHI greater than or equal to 10 were diagnosed as OSA positive. SaO_2 and HR signals from oximetry were simultaneously recorded using a dual wavelength-based finger probe with a sampling frequency of 0.2 Hz (one sample every 5 s).

We randomly divided the population under study into a training set (74 subjects, 44 patients with a positive diagnosis of OSA and 30 with a negative diagnosis of OSA) and a test set (113 subjects, 67 patients with a positive diagnosis of OSA and 46 with a negative diagnosis of OSA). The training set was used to perform the methodology and to select the optimum decision threshold. Finally, *cross-ApEn* was applied without further alteration to the data from the test set. Table 1 summarizes the demographic and clinical features of the patient groups.

Cross Approximate Entropy

The c*ross-ApEn* is a two parameter family of statistics introduced as a measure of asynchrony between two paired time series (Pincus, 2000). Its definition was proposed by Pincus as a method to compare correlated sequences, suggesting its application to physiological signals like HR and respiratory rate (Pincus & Singer, 1996).

Cross-ApEn is generally applied to compare sequences from two distinct yet intertwined variables in a network. It allows the assessment of system evolution

Table 1. Demographic and clinical characteristics of the patients groups

	ALL SUBJECTS		
	All Subjects	*OSA Positive*	*OSA Negative*
Subjects (*n*)	187	111	76
Age (*years*)	57.97 ± 12.84	58.30 ± 12.88	57.57 ± 12.87
Males (%)	78.61	84.68	69.74
BMI (*kg/m*^{2})	29.54 ± 5.51	30.45 ± 4.92	28.42 ± 6.02
Recording Time(*h*)	8.19 ± 0.62	8.17 ± 0.75	8.22 ± 0.33
AHI (n/h)		40.07 ± 19.64	2.04 ± 2.36
	TRAINING SET		
	All subjects	*OSA Positive*	*OSA Negative*
Subjects (*n*)	74	44	30
Age (*years*)	58.25 ± 12.14	56.73 ± 13.61	59.59 ± 10.19
Males (%)	75.68	79.55	70.00
BMI (*kg/m*^{2})	29.62 ± 5.71	30.19 ± 5.09	28.93 ± 6.40
Recording Time(*h*)	8.22 ± 0.41	8.20 ± 0.49	8.25 ± 0.27
AHI (*n/h*)		38.11 ± 18.18	2.60 ± 2.51
	TEST SET		
	All subjects	*OSA Positive*	*OSA Negative*
Subjects (*n*)	113	67	46
Age (*years*)	57.91 ± 13.39	59.37 ± 12.38	56.03 ± 14.54
Males (%)	80.53	88.06	69.57
BMI (*kg/m*^{2})	29.49 ± 5.41	30.63 ± 4.84	28.07 ± 5.80
Recording Time(*h*)	8.17 ± 0.72	8.14 ± 0.88	8.20 ± 0.37
AHI (*n/h*)		41.36 ± 20.58	1.67 ± 2.21

Note. Data are presented as mean ± SD or n (%). OSA Positive: patients with a positive diagnosis of obstructive sleep apnea syndrome; OSA Negative: subjects with a negative diagnosis of obstructive sleep apnea syndrome; BMI: body mass index; AHI (n/h): apnea/hypopnea index calculated for hourly periods.

characteristics such as feedback and control avoiding the requirement to model the underlying system (Pincus, 2000). *Cross-ApEn* evaluates subordinate as well as dominant patterns in data, quantifying changes in underlying episodic behavior that do not reflect in peak occurrences and amplitudes (Veldhuis et al., 1999). For two paired time series $u(i)$ and $v(i)$, *cross-ApEn* measures, within tolerance r, the (conditional) regularity or frequency of v-patterns similar to a given u-pattern of window length m. Greater asynchrony indicates fewer instances of (sub)pattern matches, quantified by larger *cross-ApEn* values (Pincus, 2000).

Formally, given two length-N sequences $u = [u(1), u(2), \ldots, u(N)]$ and $v = [v(1), v(2), \ldots, v(N)]$, and fixed the input parameters m and r, *cross-ApEn* could be computed as follows (Pincus, 2000):

1. Form the vector sequences $x(i) = [u(i), u(i+1), \ldots, u(i+m-1)]$ and $y(j) = [v(j), v(j+1), \ldots, v(j+m-1)]$. These vectors represent m consecutive u and v values respectively, commencing with the ith point.
2. Define the distance between $x(i)$ and $y(j)$, $d[x(i), y(j)]$, as the maximum absolute difference in their respective scalar components:

$$d\left[x(i), y(j)\right] = \max_{k=1,2,\ldots,m} |u(i+k-1) - v(j+k-1)| \quad (1)$$

3. For each $x(i)$, count the number of j ($j = 1\ldots N-m+1$) so that $d[x(i),y(j)] \le r$, denoted as $N_i^m(r)$. Then, for $i=1\ldots N-m+1$, set

$$C_i^m(r)(v \,\|\, u) = N_i^m(r)/(N-m+1) \quad (2)$$

The $C_i^m(r)$ values measure within a tolerance r the regularity, or frequency, of (v–) patterns similar to a given (u–) pattern of window length m.

4. Compute the natural logarithm of each $C_i^m(r)$, and average it over i,

$$\phi^m(r)(v\|u) = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} \ln C_i^m(r)(v\|u) \quad (3)$$

5. The cross approximate entropy is defined by:

$$cross\text{-}ApEn(m,r,N)(v\|u) = \phi^m(r)(v\|u) - \phi^{m+1}(r)(v\|u) \quad (4)$$

Typically, *cross-ApEn* is applied to normalized $\{u^*(i), v^*(i)\}$ time series, where $u^*(i) = [u(i)\text{–mean}(u)]/\text{SD}(u)$ and $v^*(i) = [v(i)\text{–mean}(v)]/\text{SD}(v)$.

Two important issues must be taken into account when computing *cross-ApEn* (Richman & Moorman, 2000):

a. *Cross-ApEn*$(m,r,N)(v||u)$ will not be always defined, because of $C_i^m(r)(v\|u)$ in equation (3) are equal to 0 in the absence of similar patterns between u and v. To solve this, we used two correction strategies proposed by Richman & Moorman (2000): bias 0 and bias max. Both strategies assign non zero values to $C_i^m(r)(v\|u)$ and $C_i^{m+1}(r)(v\|u)$ in the absence of any matches. The bias 0 strategy, sets the bias towards a *cross-ApEn* value of zero by setting

$$C_i^m(r) = C_i^{m+1}(r) = 1 \quad \text{if originally} \quad C_i^m(r) = C_i^{m+1}(r) = 0 \quad (5)$$
$$C_i^{m+1}(r) = (N-m)^{-1} \text{ if originally } C_i^m(r) \ne 0 \text{ and } C_i^{m+1}(r) = 0.$$

The bias max correction sets the bias towards the highest observable value of *cross-ApEn* by setting

$$C_i^m(r) = 1 \text{ if originally } C_i^m(r) = 0 \quad (6)$$
$$C_i^{m+1}(r) = (N-m+1)^{-1} \quad \text{if originally} \quad C_i^{m+1}(r) = 0$$

In this research, both correction strategies were implemented, and their influence in the diagnostic ability of *cross-ApEn* were studied in the training set.

b. There is a direction dependence. Because of the logarithm in equation (3), $\phi^m(r)(v\|u)$ will not generally be equal to $\phi^m(r)(u\|v)$. Thus, *cross-ApEn*$(v||u)$ and its direction conjugate *cross-*

ApEn($u||v$) are unequal in most cases. In the present study, both *cross-ApEn*($v||u$) and *cross-ApEn*($u||v$) were estimated, and the influence of the direction dependence in the diagnostic accuracy of *cross-ApEn* was tested in the training set.

In this research, we divided SaO_2 and HR signals in 200 sample epochs before *cross-ApEn* was applied. The input parameters were fixed to m = 1 or 2, and r was varied from 0.1 to 1.0 in steps of 0.1.

Results

Training Set. *Cross-ApEn* was applied to both OSA positive and OSA negative groups in the training set. Different combinations of the input parameters m and r were assessed, as well as the correction strategies bias 0 and bias max and the two possible directions (SaO_2||HR) and (HR||SaO_2). Student t-test was applied to check for significant differences. Firstly, a great difference was found between the directions (SaO_2||HR) and (HR||SaO_2). *Cross-ApEn*(m,r,N)(HR||SaO_2) achieved lower accuracies with all combinations of m and r in terms of the p-value. This fits well with the nature of SaO_2 and HR signals. SaO_2 signals are more regular than HR in both OSA positive and OSA negative subjects. The number of SaO_2 patterns similar to a fixed HR pattern when computing $C_i^m(r)(SaO_2\|HR)$ will be greater than the number of similar patterns when comparing a HR record to a fixed SaO_2 pattern. Hence, *cross-ApEn* will be statistically better estimated in the first case (Pincus & Goldberger, 1994), increasing the differences between OSA positive and OSA negative groups. With respect to the correction strategies, no significant differences were found between bias 0 and bias max ($p >> 0.01$).

Figure 1 displays mean *cross-ApEn*(m,r,N)(SaO_2||HR) values with bias 0 correction for both the OSA positive and the OSA negative groups in the training set. All the suggested combinations of the input parameters m and r were computed. For a fixed m and r pair, the OSA negative group achieves higher values of *cross-ApEn* than the OSA positive group. Hence, we could say that OSA negative SaO_2 and HR paired records are more asynchronous than OSA positive oximetric recordings. This behavior demonstrates the relative consistency of *cross-ApEn* applied to oximetry signals with r varying from 0.1 to 1.0 and m equal to 1 or 2.

Finally, an optimum decision threshold was derived from the training set by means of receiver operating characteristics (ROC) plots. The ROC curve is obtained representing sensitivity vs. 1-specificity for different thresholds, depending on the nature of the values provided by the method. The sensitivity of a test is the percentage of patients in the OSA positive group correctly diagnosed, whereas the specificity is the percentage of subjects in the OSA negative group rightly classified by the test. The optimum decision threshold was selected looking for a compromise between sensitivity and specificity. This value is graphically determined from

Figure 1. Mean cross-ApEn (SaO_2||HR) for both the OSA positive and OSA negative groups in the training set varying cross-ApEn input parameters

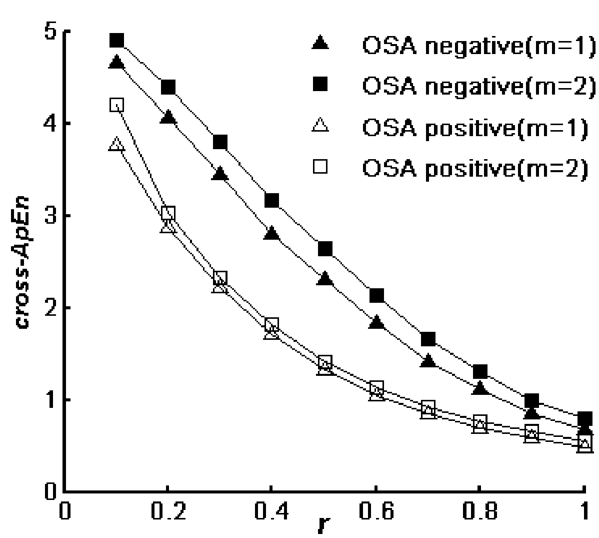


Table 2. Results from the training set with the optimum decision threshold from ROC analyses

m	*r*	Th	S(%)	E(%)	A(%)	AROC
m=1	0.1	4.38	75.0	83.3	78.4	0.87
	0.2	3.74	84.1	80.0	82.4	0.89
	0.3	2.93	88.6	80.0	85.1	0.88
	0.4	2.18	86.4	80.0	83.8	0.87
	0.5	1.67	86.4	80.0	83.8	0.86
	0.6	1.40	93.2	73.3	85.1	0.85
	0.7	0.98	81.8	76.7	79.7	0.82
	0.8	0.82	84.1	73.3	79.7	0.81
	0.9	0.71	88.6	66.7	79.7	0.76
	1.0	0.57	84.1	66.7	77.0	0.77
m=2	0.1	4.85	79.6	76.7	78.4	0.84
	0.2	3.71	77.3	86.7	81.1	0.89
	0.3	3.29	88.6	80.0	85.1	0.89
	0.4	2.51	86.4	80.0	83.8	0.88
	0.5	1.65	81.8	83.3	82.4	0.87
	0.6	1.70	95.5	73.3	86.5	0.87
	0.7	1.10	81.8	76.7	79.7	0.83
	0.8	0.91	84.1	76.7	81.1	0.83
	0.9	0.79	90.9	66.7	81.1	0.79
	1.0	0.63	81.8	70.0	77.0	0.79

Note. Th: Optimum threshold; S: Sensitivity; E: Specificity; A: Accuracy; AROC: Area under the ROC curve.

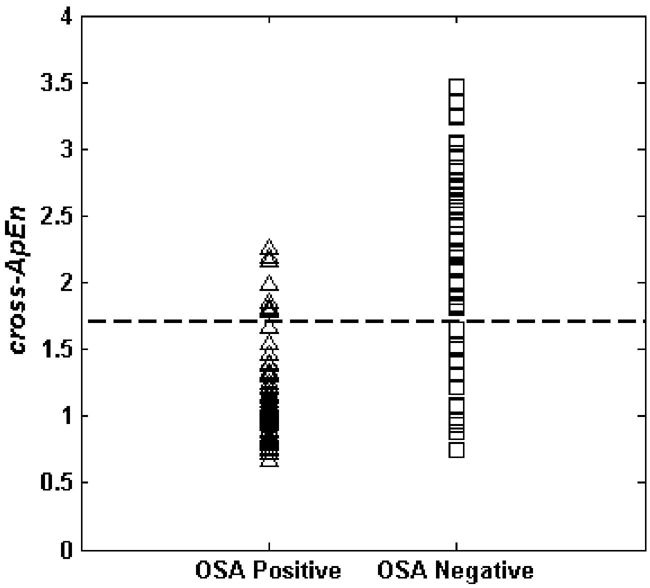
the ROC curve as the closest point to the left top corner (100% sensitivity and 100% specificity). Table 2 shows the statistics for the optimum threshold from the ROC analysis for *cross-ApEn*(*m*,*r*,*N*)(SaO_2||HR) with bias 0 correction. The best diagnostic accuracy was achieved with m = 2 and r = 0.6. The optimum threshold was selected at 1.7, where we achieved 95.5% sensitivity, 73.3% specificity, 86.5% accuracy and an area under the ROC curve of 0.87.

Test Set. We applied *cross-ApEn* to the test set using the optimum parameters derived from the training set: m = 2, r = 0.6, (SaO_2||HR) direction and bias 0 correction. Figure 2 displays the mean *cross-ApEn* values for the OSA positive and the OSA negative subjects in the test set. The dashed line represents the optimum decision threshold derived from the training set. Patients in the OSA positive group were characterized by lower *cross-ApEn* values (1.16 ± 0.38), according to the greater synchrony between their SaO_2 and HR recordings. On the other hand, OSA negative subjects were characterized by higher *cross-ApEn* values (2.18 ± 0.73). Finally, the diagnostic ability of *cross-ApEn* in screening for OSA syndrome was assessed computing the sensitivity, specificity and accuracy values in the test set with the optimum decision threshold provided by the training set. We achieved 86.7% sensitivity, 73.9% specificity, and 81.4% accuracy.

Discussion

In the present study, we analyzed the oximetric recordings of a sample of 187 patients (111 with a positive diagnosis of OSA and 76 with a negative diagnosis of OSA) looking for an alternative method to PSG in the diagnosis of OSA. The whole data set were randomly divided in a training set (44 OSA positive patients and 30 OSA negative subjects) and a test set (67 OSA positive patients and 46 OSA negative subjects). The training set was used to define different methodological issues and to select an optimum decision threshold, whereas the subsequently use of the test set provided a higher generalization of the results. A nonlinear method called *cross-ApEn* was applied to paired SaO_2

Figure 2. Mean cross-ApEn (SaO_2||HR) values from nocturnal oximetry for patients in the OSA positive and the OSA negative group in the test set. The horizontal dashed line, set at 1.7034, represents the optimum decision threshold between both populations derived from the training set.



and HR signals from nocturnal oximetry to quantify their asynchrony. A shorter value of c*ross-ApEn* indicates greater instances of (sub)pattern matches, so the sequences are more synchronous (Pincus, 2000). We found that the recurrent *apnea events* in OSA positive patients lead to a significant decrease in the c*ross-ApEn* values. Oxygen desaturations associated with apnea events typical of OSA, cause coordinate fluctuations in both SaO_2 and HR signals, leading to low c*ross-ApEn* values. Furthermore, we found a great direction dependence of the method. *Cross-ApEn*(SaO_2||HR) is better estimated than *cross-ApEn*(HR||SaO_2) ($p < 0.01$). On the other hand, no significant differences were found between the correction strategies bias 0 and bias max ($p >> 0.01$).

Previous studies assessed the relationship between oscillations in SaO_2 and HR as a sign of periodic breathing (Keyl, Lemberger, Pfeifer, Hochmuth, & Gesiler, 1997; Zamarrón et al., 2005). The study developed by Zamarrón et al. (2005) showed that the average level of SaO_2 and HR coherence in OSA positive patients was significantly higher than in OSA negative patients. Their results fit well with those obtained in our study, indicating higher synchrony between oximetry signals from OSA subjects. A sensitivity of 86.6%, a specificity of 73.9%, and an accuracy of 81.4% were obtained with the test set, using the optimum threshold of 1.7 derived from the training set.

We should take into account some limitations of our study. The sample size could be larger and OSA positive patients were predominantly studied. Moreover, a drawback in the applicability of our method should be mentioned. Oximetry signals were recorded simultaneously with PSG, eliminating potential confounders such as night to night variability of AHI, as well as ensuring that oximetry data were collected in exactly the same environment as the PSG data. Further analyses using unattended nocturnal oximetry in home are necessary. An additional difficulty lies in the estimation of the HR from the distal pulse obtained from pulse oximetry, instead of measuring the elapsed time between consecutive R waves. An overestimation of the high frequencies must be taken into account.

FUTURE TRENDS

Last years has shown an increasing interest in assessing portable monitoring as an alternative to PSG to study patients with suspicion of OSA syndrome. Its simplicity and low cost have made nocturnal oximetry an excellent screening method previous to PSG. However, research studies typically record nocturnal oximetry and PSG simultaneously in a sleep laboratory. Hence, the cost effectiveness of unattended monitoring at home would need to be evaluated prospectively in a controlled trial (Whitelaw et al., 2005). Moreover, additional work is required to apply our methodology to larger data set with a wide spectrum of sleep-related breathing disorders, as well as to study groups of especial interest, such as patients with lung diseases and young snorers.

Several quantitative indexes derived from nocturnal oximetric recordings are widely used by physicians, especially ODIs and CT. However, there is not a standardized definition or consensus on its use. Our research group has recently assessed the usefulness of nonlinear analysis in the diagnosis of OSA syndrome. We have shown that regularity, variability, and complexity measures from SaO_2 could improve the diagnostic accuracy of classical oximetric indexes (Hornero et al., 2007; Álvarez et al., 2006). Moreover, in the present study, we have found that HR could provide additional information to SaO_2 using *cross-ApEn*. However, it would be necessary to assess additional nonlinear methods, mainly those oriented to analyze the relationship between various signals. Further work is needed to assess nonlinear analysis and to verify its advantages over classical oximetric indexes and the linear approaches.

CONCLUSION

In summary, we found that the recurrence of apnea events in patients with OSA determined a significant decrease in the *cross-ApEn* values of the SaO_2 and HR oximetric recordings. This suggests that the apnea events cause synchrony fluctuations in SaO_2 and HR, affecting both respiratory and circulatory systems and their relationship. A sensitivity of 86.6%, specificity of 73.9%, and accuracy of 81.4% were achieved. With this research, we verified that the analysis of SaO_2 and HR jointly by means of *cross-ApEn* could provide useful information to help in the OSA diagnosis.

REFERENCES

Álvarez, D., Hornero, R., Abásolo, D., del Campo, F., & Zamarrón, C. (2006). Nonlinear characteristics of blood oxygen saturation from nocturnal oximetry for obstructive sleep apnoea detection. *Physiological Measurement, 27,* 399-412.

Barbé, F., Pericas, J., Muñoz, A., Findley, L., Anto, J.M., & Agustí, A. (1998). Automobile accidents in patients with sleep apnea syndrome. An epidemiological study. *American Journal of Respiratory and Critical Care Medicine, 158*, 18-22.

Chaudhary, B., Dasti, S., Park, Y., Brown, T., Davis, H., & Akhtar, B. (1998). Hour-to-hour variability of oxygen saturation in sleep apnea. *Chest, 113,* 719-722.

Douglas, N.J. (2002). Recent advances in the obstructive sleep apnoea/hypopnoea syndrome. *Annals Academy of Medicine, 31,* 697-701.

Flemons, W.W., & Littner, M.R. (2003). Measuring agreement between diagnostic devices. *Chest, 124,* 1535-–1542.

Gami, A.S., Howard, D.E., Olson, E.J., & Somers, V.K. (2005). Day–night pattern of sudden death in obstructive sleep apnea. *The New England Journal of Medicine, 352*, 1206-1214.

Hampson, M.H., & Mallen, E.A.H. (2006). Coherence function analysis of the high-order aberrations of the human eye. *Optics Letters, 31*, 184-186.

Hornero, R., Álvarez, D., Abásolo, D., del Campo, F., & Zamarrón, C. (2007). Utility of approximate entropy from overnight pulse oximetry data in the diagnosis of the obstructive sleep apnea syndrome. *IEEE Transactions on Biomedical Engineering, 54,* 107-113.

Keyl, C., Lemberger, P., Pfeifer, M., Hochmuth, K., & Gesiler, P. (1997). Heart rate variability in patients with daytime sleepiness suspected of having sleep apnea syndrome: A receiver-operating characteristics analysis. *Clinical Sciences, 92*, 335-343.

Leung, R.S.T., & Bradley, T.D. (2001). Sleep apnea and cardiovascular disease. *American Journal of Respiratory and Critical Care Medicine, 164*, 2147-2165.

Liu, P.Y., Pincus, S.M., Keenan, D.M., Roelfsema, F., & Veldhuis, J.D. (2005). Analysis of bidirectional pattern synchrony of concentration-secretion pairs: Implementation in the human testicular and adrenal axes. *American Journal of Physiology—Regulatory, Integrative and Comparative Physiology, 288,* R440-R446.

Netzer, N., Eliasson, A.H., Netzer, C., & Kristo, D.A. (2001). Overnight pulse oximetry for sleep-disordered breathing in adults. *Chest, 120,* 625-633.

Nunez, P.L., Srinivasan, R., Westdorp, A.F., Wijesinghe, R.S., Tucker, D.M., Silberstein, R.B., et al. (1997). EEG coherency I: Statistics, reference electrode, volume conduction, Laplacians, cortical imaging, and interpretation at multiple scales. *Electroencephalography and Clinical Neurophysiology, 103*, 499-515.

Pincus, S.M. (2000). Irregularity and asynchrony in biologic network signals. *Methods in Enzymology, 321*, 149-182.

Pincus, S.M., & Goldberger, A.L. (1994). Physiological time-series analysis: What does regularity quantify? *American Journal of Physiology, 266*, H1643-H1656.

Pincus, S.M., Mulligan, T., Iranmanesh, A., Gheorghiu, S., Godschalk, M., & Veldhuis, J.D. (1996). Older males secrete luteinizing hormone and testosterone more irregularly, and jointly more asynchronously, than younger males. In *Proceedings of the National Academy of Sciences of the United States of America* (Vol. 93, pp. 14100-14105).

Pincus, S.M., & Singer, B.H. (1996). Randomness and degrees of irregularity. In *Proceedings of the National Academy of Sciences of the United States of America* (Vol. 93, pp. 2083-2088).

Polysomnography Task Force. American Sleep Disorders Association Standards of Practice Committee. (1997). Practice parameters for the indications of polysomnography and related procedures. *Sleep, 20,* 406-422.

Porta, A., Baselli, G., Lombardi, F., Montano, N., Malliani, A., & Cerutti, S. (1999). Conditional entropy approach for the evaluation of coupling strength. *Biological Cibernetics, 81,* 119-129.

Porta, A., Guzzetti, S., Montano, N., Pagani, M., Somers, V., Malliani, A., et al. (2000). Information domain analysis of cardiovascular variability signals: Evaluation of regularity, synchronization and co-ordination. *Medical & Biological Engineering & Computing, 38,* 180-188.

Rechtschaffen, A., & Kales, A. (1968). *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects*. Los Angeles, CA: Brain Information Services, Brain Research Institute, University of California.

Richman, J.S., & Moorman, J.R. (2000). Physiological time series analysis using approximate entropy and sample entropy. *American Journal of Physiology Heart and Circulatory Physiology, 278,* H2039-H2049.

Schwartz, D.J., Kohler, W.C., & Karatinos, G. (2005). Symptoms of depression in individuals with obstructive sleep apnea may be amenable to treatment with continuous positive airway pressure. *Chest, 128,* 1304-1309.

Stam, C.J., & van Dijk, B.W. (2002). Synchronization likelihood: An unbiased measure of generalized synchronization in multivariate data sets. *Physica D, 163*, 236-251.

Veldhuis, J.D., Iranmanesh, A., Mulligan, T., & Pincus, S.M. (1999). Disruption of the young-adult synchrony between luteinizing hormone release and oscillations in follicle-stimulating hormone, prolactin, and nocturnal penile tumescence (NPT) in healthy older men. *The Journal of Clinical Endocrinology and Metabolism, 84*, 3498-3505.

Veldhuis, J.D., Pincus, S.M., García-Rudaz, M.C., Ropelato, M.G., Escobar, M.E., & Barontini, M. (2001). Disruption of the joint synchrony of luteinizing hormone, testosterone, and androstenedione secretion in adolescents with polycystic ovarian syndrome. *The Journal of Clinical Endocrinology and Metabolism, 86,* 72-79.

Whitelaw, W.A., Brant, R.F., & Flemons W.W. (2005). Clinical usefulness of home oximetry compared with polysomnography for assessment of sleep apnea. *American Journal of Respiratory and Critical Care Medicine, 171,* 188-193.

Young, T., Peppard, P.E., & Gottlieb, D.J. (2002). Epidemiology of obstructive sleep apnea. *American Journal of Respiratory and Critical Care Medicine, 165,* 1217-1239.

Zamarrón, C., Gude, F., Barcala, J., Rodríguez, J.R., & Romero, P.V. (2003). Utility of oxygen saturation and heart rate spectral analysis obtained from pulse oximetric recordings in the diagnosis of sleep apnea syndrome. *Chest*, *123***,** 1567-1576.

Zamarrón, C., Romero, P.V., Gude, F., Amaro, A., & Rodríguez, J.R. (2001). Screening of obstructive sleep apnea: Heart rate spectral analysis of pulse oximetric recording. *Respiratory Medicine*, *95*, 759-765.

Zamarrón, C., Pichel, F., & Romero, P.V. (2005). Coherence between oxygen saturation and heart rate obtained from pulse oximetric recordings in obstructive sleep apnoea. *Physiological Measurement*, *26*, 799-810.

KEY TERMS

Apnea Event: The omplete cessation of airflow for at least 10 seconds.

Blood Oxygen Saturation (SaO_2): A percentage of hemoglobin binding sites in the bloodstream occupied by oxygen.

Cross Approximate Entropy (Cross-ApEn): The nonlinear measure of asynchrony between two paired time series.

Heart Rate (HR): The frequency of the cardiac cycle equal to the number of contractions of the heart (heart beats) in 1 minute commonly expressed as beats per minute (bpm).

Hypopnea Event: The partial reduction of respiratory flow for at least 10 seconds accompanied by a 4% or more decrease in the saturation of hemoglobin.

Nonlinear System: A system whose behavior is not subject to the principle of superposition, that is, its behavior is not expressible as the sum of its parts.

Obstructive Sleep Apnea (OSA) Syndrome: A sleep disorder characterized by repetitive obstruction of the upper airway despite the presence of respiratory effort, resulting in oxygen desaturations and arousals during sleep.

Oximetry: The noninvasive indirect measure of the percentage of arterial oxygen combined with the hemoglobin blood oxygen saturation (SaO_2), usually accompanied by an estimation of the heart rate (HR).

Polysomnography (PSG): A multiparametric test used in the study of sleep and breathing usually performed during the night in a specialized sleep unit. PSG records biophysiological and neurophysiological data, including electroencephalogram (EEG), electrocardiogram (ECG), electrooculogram (EOG), electromiogram (EMG), measurement of chest wall movement and airflow, and oximetry.

Applying Independent Component Analysis to the Artifact Detection Problem in Magnetoencephalogram Background Recordings

Javier Escudero
University of Valladolid, Spain

Roberto Hornero
University of Valladolid, Spain

Daniel Abásolo
University of Valladolid, Spain

Jesús Poza
University of Valladolid, Spain

Alberto Fernández
Complutense University of Madrid, Spain

INTRODUCTION

The analysis of the electromagnetic brain activity can provide important information to help in the diagnosis of several mental diseases. Both electroencephalogram (EEG) and magnetoencephalogram (MEG) record the neural activity with high temporal resolution (Hämäläinen, Hari, Ilmoniemi, Knuutila, & Lounasmaa, 1993). Nevertheless, MEG offers some advantages over EEG. For example, in contrast to EEG, MEG does not depend on any reference point. Moreover, the magnetic fields are less distorted than the electric ones by the skull and the scalp (Hämäläinen et al., 1993). Despite these advantages, the use of MEG data involves some problems. One of the most important difficulties is that MEG recordings may be severely contaminated by additive external noise due to the intrinsic weakness of the brain magnetic fields. Hence, MEG must be recorded in magnetically shielded rooms with low-noise SQUID (Superconducting QUantum Interference Devices) gradiometers (Hämäläinen et al., 1993).

Unfortunately, the external noise is not the only undesired signal in MEG data. In these recordings, noncerebral sources (i.e., artifacts) appear mixed with the useful brain signals. The artifacts could bias the brain activity analyses, since both kinds of signals may have similar power and share the same frequency band. In MEG data, the main artifact is the cardiac one, whose amplitude is usually high enough to be visible in raw recordings (Jousmäki & Hari, 1996). Similarly, the ocular artifacts can be evident in MEG data (Antervo, Hari, Katila, Ryhänen, & Seppänen, 1985), and they can disguise the actual brain activity in long recordings. Finally, power line noise may also be present in MEG signals (Hämäläinen et al., 1993).

Diverse approaches have been used to detect and reject artifacts from EEG and MEG data, such as epoch rejection, regression methods (Croft & Barry, 2000), or principal component analysis (PCA) (Sadasivan & Dutt, 1996). In the early 1990s, a new method to obtain a blind source separation (BSS) became available: the independent component analysis (ICA) (Comon, 1994; Jutten & Herault, 1991). Since then, ICA has been increasingly used in the artifact rejection problem (Delorme, Makeig, & Sejnowski, 2001; Escudero, Hornero, Abásolo, Poza, Fernández, & López, 2006; James & Hesse, 2005; Jung, Makeig, Humphries, Lee, McKeown, Iragui, & Sejnowski, 2000; Sander, Wübbeler, Lueschow, Curio, & Trahms, 2002; Vigário, 1997; Vigário, Jousmäki, Hämäläinen, Hari, & Oja, 1998). One of the main advantages of ICA over other approaches is that artifacts must not be orthogonal to brain signals, and reference channels are not needed, although they can help to detect the artifacts (Barbati, Porcaro, Zappasodi, Rossini, & Tecchio, 2004; Flexer, Bauer, Pripfl, & Dorffner, 2005; Joyce, Gorodnitsky, & Kutas, 2004).

BACKGROUND

The use of ICA in the artifact rejection problem can be summarized as follows. Firstly, ICA finds a separating matrix, which decomposes the multidimensional input data (the EEG or MEG channels) into several independent components (ICs). Next, the ICs which account for artifacts are marked by visual inspection (Vigário, 1997; Vigário et al., 1998) or automatic methods (Barbati et al., 2004; Escudero et al., 2006; Flexer et al., 2005; Joyce et al., 2004; Li, Ma, Lu, & Li, 2006; Ting, Fung, Chang, & Chan, 2006), and they are removed. Finally, the pseudo-inverse of the separating matrix is used to build the output data (i.e., the signals without artifacts) from the retained ICs.

In this process, the main open issue is the artifact recognition. Several automatic criteria have been proposed. For instance, Delorme et al. (2001) used kurtosis and entropy to identify artifacts in EEG data. Adding a correlation criterion to these parameters, Barbati et al. (2004) proposed a semiautomatic approach to detect various artifacts in MEG signals. Similarly, Joyce et al. (2004) used a correlation metric, together with other simple procedures to remove ocular artifacts from EEG data. Recently, three criteria based on amplitude thresholds, power in frequency bands, and scalp distributions of the ICs were used by Ting et al. (2006) to deal with muscular and ocular artifacts in EEG. Moreover, Escudero et al. (2006) showed that a criterion based on the skewness of the IC amplitude distributions could detect the cardiac-related ICs better than kurtosis in MEG data.

In addition, it is important to select the number of ICs correctly. Whereas some papers set a power threshold (Ting et al., 2006), other studies set this parameter to the number of available channels (Flexer et al., 2005; Joyce et al., 2004; Jung et al., 2000; Li et al., 2006; Sander et al., 2002). Nevertheless, few statistical criteria have been used (Cao, Murata, Amari, Cichocki, & Takeda, 2003; Escudero et al., 2006; Ikeda & Toyama, 2000).

As it can be noted, very different approaches have been proposed to remove artifacts from EEG and MEG data with ICA. However, none has been generally adopted. This could be due to the intrinsic complexity of the MEG signals and the ICA algorithms. Because of this reason, straightforward artifact detection criteria might be preferred to complex ones, in order to keep the whole artifact rejection process at a suitable level of complexity. Therefore, in this study, we evaluated the usefulness of three criteria based on higher order statistics and spectral properties to detect the cardiac, ocular, and the power line artifacts in MEG data.

MAGNETOENCEPHALOGRAM RECORDING

Seven elderly subjects without past or present neurological disorders—age = 65.6 ± 7.9 years; mean ± standard deviation (SD)—participated in this study. They were asked to stay awake with closed eyes, and to reduce eye and head movements while they were lying on a patient bed to record the MEG signals. These conditions are similar to the recording protocol used in diagnostic studies. All subjects gave their informed consent for the participation in this study, which was approved by the local ethics committee.

For each subject, five minutes of MEG data were acquired at a sampling frequency of 678.17 Hz with a 148-channel whole-head magnetometer (MAGNES 2500 WH, 4D Neuroimaging) in a magnetically shielded room. The recordings were downsampled to 169.549 Hz. Fourteen epochs of 50 s (8,477 samples) with artifacts were selected for off-line analysis. All of them had cardiac artifacts, and some also presented ocular and/or power line ones. The epochs were digitally filtered using a band-pass filter with cut-off frequencies at 0.5 Hz and 60 Hz.

ARTIFACT REJECTION METHOD

Independent Component Analysis

The ICA model assumes that the n MEG channels, $\mathbf{x}(t) = [x_1(t), \ldots, x_n(t)]^T$, are a linear mixture of m ICs, $\mathbf{s}(t) = [s_1(t), \ldots, s_m(t)]^T$, with $m \leq n$ (James & Hesse, 2005). In order to represent the external additive noise, an n-dimensional vector of spatially uncorrelated Gaussian noise, $\mathbf{v}(t)$, can also be included in the model (Barbati et al., 2004; Cao et al., 2003; Ikeda & Toyama, 2000; Ting et al., 2006):

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{v}(t) \quad (1)$$

where the means of $\mathbf{s}(t)$, $\mathbf{x}(t)$, and $\mathbf{v}(t)$ are zero and $\mathbf{A}$ is a full rank $n \times m$ mixing matrix.

In this problem, only $\mathbf{x}(t)$ is known. Therefore, $\mathbf{A}$, $\mathbf{s}(t)$, and $\mathbf{\Psi}$—the covariance matrix of the external noise—have to be estimated blindly from $\mathbf{x}(t)$. To achieve this BSS, ICA makes several assumptions about the data (Cichocki & Amari, 2002; Comon, 1994): the ICs are statistically independent and have non-Gaussian distributions, and the mixing process is linear, instantaneous, and stationary. All these hypotheses have been previously validated for EEG and MEG data (James & Hesse, 2005; Vigário & Oja, 2000).

ICA provides a separating matrix, $\mathbf{W} = \mathbf{A}^{+}$ (apex "+" denotes a pseudo-inverse matrix). Using $\mathbf{W}$, we can compute:

$$\mathbf{y}(t) = \mathbf{W}\mathbf{x}(t) = \mathbf{W}[\mathbf{A}\mathbf{s}(t) + \mathbf{v}(t)] = \mathbf{s}(t) + \mathbf{W}\mathbf{v}(t) \quad (2)$$

where $\mathbf{y}(t) = [y_1(t), \ldots, y_m(t)]^{\mathrm{T}}$ is the m-dimensional vector which estimates $\mathbf{s}(t)$ (i.e., the ICs).

To calculate $\mathbf{W}$, we have used the cumulant-based iterative inversion (CII) algorithm (Cruces-Alvarez, Cichocki, & Castedo-Ribas, 2000). This ICA algorithm is robust to external Gaussian noise, in the sense that it is able to find an asymptotically unbiased estimation of $\mathbf{W}$, when calculations are carried out with enough samples (typically ≥ 5000) (Barbati et al., 2004; Cruces-Alvarez et al., 2000; Escudero et al., 2006).

Preprocessing

When high-density recording equipment is used, it may be necessary to reduce the data dimensionality (James & Hesse, 2005; Vigário & Oja, 2000). However, there is no information about the real m value. This dimensionality reduction is usually carried out with PCA (James & Hesse, 2005; Vigário & Oja, 2000), but this approach has some drawbacks, since it assumes that all the MEG channels have the same noise power (Barbati et al., 2004; Ting et al., 2006), and it does not set an objective criterion to estimate m (James & Hesse, 2005). To avoid these drawbacks, we have implemented the preprocessing using the unweighted least squares method (Cao et al., 2003) of factor analysis (FA) (Reyment & Jöreskog, 1993), which considers $\mathbf{\Psi}$. The model for this preprocessing is:

$$\mathbf{x}(t) = \mathbf{A}_{\mathrm{Pr}}\mathbf{z}(t) + \mathbf{v}(t) \quad (3)$$

where $\mathbf{C}$ is the covariance matrix of $\mathbf{x}(t)$, and $\mathbf{A}_{\mathrm{Pr}}$ as the preprocessing mixing matrix that relates the preprocessed data, $\mathbf{z}(t)$, to $\mathbf{x}(t)$.

$\mathbf{A}_{\mathrm{Pr}}$ and $\mathbf{\Psi}$ can be estimated from $\mathbf{C}$ using an iterative procedure (Cao et al., 2003, Escudero et al., 2006). Then, the matrix $\mathbf{Q}$ used to preprocess the data is calculated as the pseudo-inverse of $\mathbf{A}_{\mathrm{Pr}}$ taking into account the external noise power (Cao et al., 2003, Escudero et al., 2006). Finally, the preprocessed data are obtained by:

$$\mathbf{z}(t) = \mathbf{Q}\mathbf{x}(t) \quad (4)$$

This procedure assumes that m is known. However, m must also be estimated from the data. In order to determine it, we used a method derived from FA and based on statistical model selection with information criteria: the minimum description length (MDL), which was computed according to (Ikeda & Toyama, 2000) (Box 1),where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix and N is the number of data points.

Artifact Detection Methods

Power Line Noise Artifacts. ICA can isolate most power line noise into one IC (Jung et al., 2000). Hence, the spectrum of such IC should be centered at the power line frequency (50 Hz in Europe). A spectral parameter, P_{50Hz}, which measures the fraction of the power spectral density (PSD) contained between 49.5 Hz and 50.5 Hz, may be useful to mark this kind of ICs. Therefore, we

Box 1.

$$MDL = \frac{1}{2}\mathrm{tr}\left[\mathbf{C}\left(\mathbf{\Psi} + \mathbf{A}_{\mathrm{pr}}\mathbf{A}_{\mathrm{pr}}^{\mathrm{T}}\right)^{-1}\right] + \frac{1}{2}\log\left[\det\left(\mathbf{\Psi} + \mathbf{A}_{\mathrm{pr}}\mathbf{A}_{\mathrm{pr}}^{\mathrm{T}}\right)\right]\cdots$$
$$\cdots + \frac{n}{2}\log(2\pi) + \frac{\log N}{N}\left[n(m+1) - \frac{m(m-1)}{2}\right]$$

marked for rejection the ICs with P_{50Hz} values larger than a threshold (th_{50Hz}).

Cardiac Artifacts. Skewness (Skw) is defined as:

$$Skw = \frac{m_3}{(m_2)^{3/2}}, \qquad (6)$$

where $\mathrm{E}\{\cdot\}$ is the expectation value, and m_n is the nth central moment of a statistical distribution: $m_n = \mathrm{E}\left\{\left(x - \mathrm{E}\{x\}\right)^n\right\}$. *Skw* measures the asymmetry degree of a distribution. It is zero only if the distribution is symmetrical. Thus, large absolute values of *Skw* should be related to asymmetrical ICs amplitude distributions, which are typical of cardiac artifacts (Escudero et al., 2006).

Ocular Artifacts. The spectrum of the ocular artifacts is usually focused in lower frequencies than in cardiac or brain activity (Joyce et al., 2004). It follows that the ICs whose energy is mainly gathered in a very low frequency band could be due to the ocular artifacts. Therefore, we calculated the fraction of the PSD, which every IC has from 0.5 Hz to 2.5 Hz (P_{LF}), and we set a threshold (th_{LF}). An IC is considered an ocular artifact when $P_{LF} > th_{LF}$.

RESULTS

Fourteen epochs of 50 s were processed to detect the cardiac, ocular, and power line artifacts. Firstly, the optimal *m* value for every MEG was estimated in the preprocessing stage. From the 148 MEG channels, the average estimated number of ICs was 30.14 ± 7.51 ICs (mean ± SD). Once the epochs had been preprocessed, we performed the ICA decomposition using the CII algorithm. The estimated ICs were visually inspected in both time and frequency domains, and they were compared to the raw MEG recordings. By this procedure, the ICs were classified into four groups: cardiac ICs, ocular ICs, power line ICs, and other signals. In all epochs, one IC was responsible for the cardiac artifact. Since the MEG was recorded with closed eyes, only 10 of the 14 analyzed MEG epochs had ocular artifacts. From these 10 epochs, 17 ICs were associated with ocular artifacts. Finally, 12 of the 14 MEG epochs had power line interference. In these epochs, the major line noise component was isolated into one IC.

Afterward, we used three straightforward metrics to detect the artifact ICs. First, we used the P_{50Hz} parameter to assess the line noise in every IC. Setting $th_{50Hz} = 0.500$, all the power line ICs were automatically recognized. Figure 1(a) illustrates this kind of IC. Actually, there was a large difference between the P_{50Hz} values for the power line ICs, which provided P_{50Hz} values larger than 0.586, and for any other IC, with values lower than 0.041, except for an IC with P_{50Hz} = 0.146.

We also used the *Skw* of the IC amplitude distributions to mark the cardiac artifacts. The IC that provides the maximum absolute value of *Skw* for every MEG epoch could be related to the cardiac artifact. We found that this criterion marked this artifact correctly in all cases. Figure 1(b) shows this artifact recognition principle. It can be noted that this artifact has an asymmetrical distribution, which produces large *Skw* values.

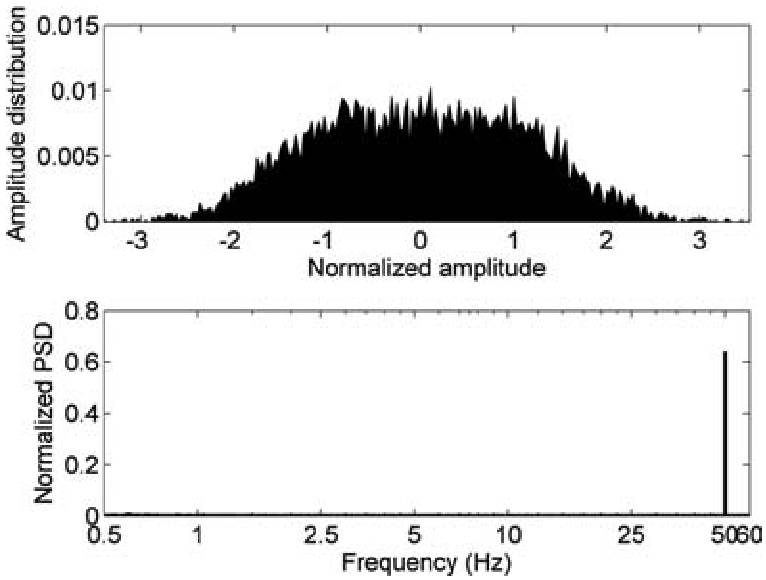
Finally, we used the spectra of the ICs to detect the ocular artifacts. Since a MEG epoch with ocular artifacts may produce more than one ocular-related IC, we evaluated the performance of the P_{LF} parameter using the 17 ocular ICs and the 17 nonocular ICs which had the largest P_{LF} values. We used this group of 34 ICs to find the optimal value for th_{LF}, which was 0.349. With this threshold, 16 of the 17 ocular ICs were correctly classified (94.12%), although three nonartifact ICs were erroneously marked as ocular artifacts. Figure 1(c) plots the spectra and amplitude distribution of an ocular IC.

DISCUSSION

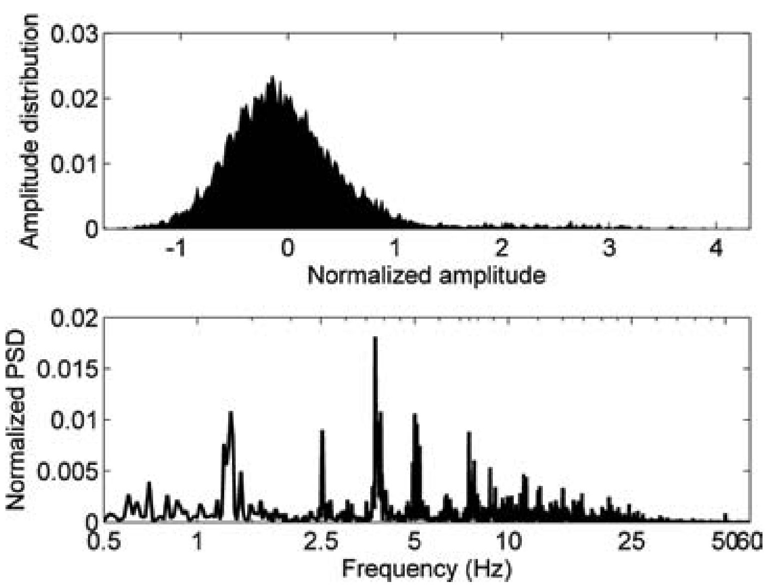
Although MEG can provide useful information about the brain, it is usually polluted by several artifacts, which could bias the analyses (Hämäläinen et al., 1993). Diverse approaches, such as epoch rejection, regression techniques (Croft & Barry, 2000), or PCA (Sadasivan & Dutt, 1996), have been traditionally used to remove the artifacts. However, these approaches present certain drawbacks (Jung et al., 2000; Vigário & Oja, 2000). ICA has been proposed as an alternative to these methods, since the hypotheses ICA makes about the data seem sensible for EEG and MEG recordings (James & Hesse, 2005).

We aimed to remove the cardiac, ocular, and the power line artifacts from MEG recordings with ICA. Firstly, we preprocessed the MEG epochs to automatically select the number of ICs for each of them (Cao et

Figure 1. Examples of different kinds of ICs extracted from the same MEG epoch. The normalized PSD and amplitude distributions of the ICs are shown. (a) Power line IC. (b) Cardiac IC. (c) Ocular IC.



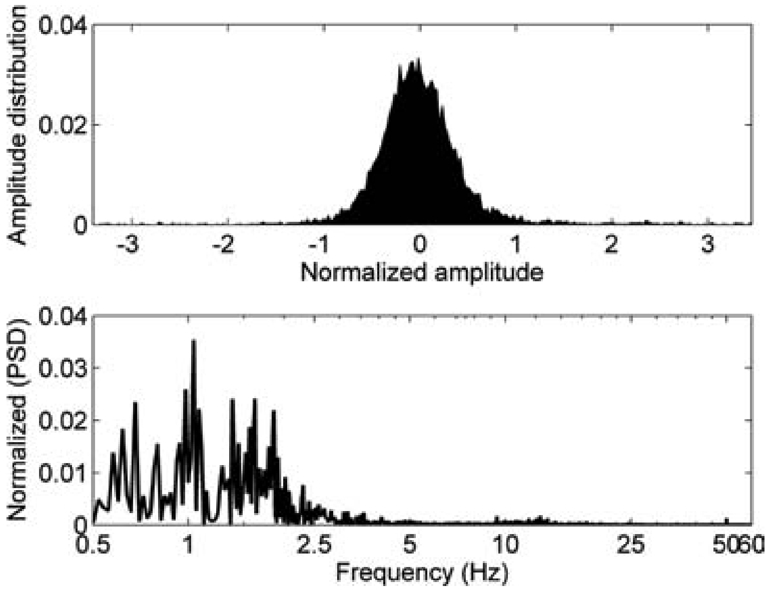
(a)



(b)

continued on following page

Figure 1. continued



(c)

al., 2003; Escudero et al., 2006; Ikeda & Toyama, 2000). The mean number of ICs was 30.14 ± 7.51. This method contrasts with other criteria which involve a certain degree of subjectivity, such as setting $n = m$ (Flexer et al., 2005; Joyce et al., 2004; Jung et al., 2000; Li et al., 2006; Sander et al., 2002) or using a power threshold (Ting et al., 2006). Afterwards, the CII algorithm was applied to estimate the ICs, which were inspected and classified. Finally, three straightforward parameters were used to detect the studied artifacts.

To identify the ICs that account for the power line interferences, we calculated the fraction of the signal PSD centered at 50 Hz. Setting a threshold $th_{50Hz} = 0.500$, we could detect the 12 power line ICs. Moreover, the largest P_{50Hz} values were always found in this kind of ICs. This confirms that ICA can isolate most power line noise into only one component (Jung et al., 2000).

The results also showed that *Skw* could detect all the 14 cardiac-related ICs, since *Skw* marks asymmetrical distributions, which are typical of the cardiac signal. On the other hand, other higher-order statistics, such as kurtosis or entropy, do not distinguish between cardiac and ocular artifacts (Barbati et al., 2004; Escudero et al., 2006).

Finally, setting a threshold on the power contained between 0.5 Hz and 2.5 Hz ($th_{LF} = 0.349$), we could identify 16 of the 17 ocular ICs (94.12%). However, three nonartifact ICs were incorrectly marked as ocular ICs. This kind of error should be avoided, as it can lead to remove part of the brain activity. Due to the high rate of ocular ICs correctly classified by P_{LF}, this criterion may be supplemented with another parameter to avoid those false artifacts, similarly to (Joyce et al., 2004), who used three criteria together to remove ocular artifacts from EEG recordings.

Certain limitations of our study merit attention. Firstly, the sample size was small, and further analyses have to be performed to confirm our findings. Moreover, evaluation of artifact detection is difficult, since the true unmixed signals are not known. Thus, it is impossible to completely validate the physiological meaning of every IC.

FUTURE TRENDS

In the last years, ICA-based approaches have been used to extract and analyze ICs from brain activity off-line. However, some applications, like brain-computer interface, need real-time algorithms. To deal with these applications, online artifact removal algorithms have been recently proposed (Shayegh & Erfanian, 2006), although they are only a first step toward this aim.

In addition, the application of ICA and, more generally, BSS techniques in brain recording analysis should not be limited to the artifact rejection. These methods might be applied to transform the complex EEG and MEG recordings into simpler components, which could provide a better understanding of the brain dynamics. For example, Jin, Jeong, Jeong, Kim, and Kim (2002) applied ICA to analyze the effect of sound and light stimulation on the ICs of the EEG. They found that ICA could improve the interpretation of the changes that occur in the EEG before and after stimulation. Moreover, Cichocki, Shishkin, Musha, Leonowicz, Asada, and Kurachi (2005) decomposed the EEG from mild cognitive impairment patients and control subjects. Their results showed that this EEG filtering—based on a BSS and a projection of the components which are more sensitive to the disease—could improve the rate of correctly classified patients and controls.

CONCLUSION

Our analysis of the ICs which composed 14 MEG epochs from seven elderly subjects suggests that all the major power line interferences could be rejected, setting a threshold on the fraction of their PSD at 50 Hz. Moreover, skewness could also detect all cardiac ICs. In addition, the fraction of the energy that an IC has between 0.5 Hz and 2.5 Hz could be a useful parameter to recognized ocular-related ICs, although this parameter may incorrectly mark nonocular ICs. Therefore, it should be supplemented with another criterion, which could avoid those false positives.

REFERENCES

Antervo, A., Hari, R., Katila, T., Ryhänen, T., & Seppänen, M. (1985). Magnetic fields produced by eye blinking. *Electroencephalography and Clinical Neurophysiology, 61*(4), 247–253.

Barbati, G., Porcaro, C., Zappasodi, F., Rossini, P. M., & Tecchio, F. (2004). Optimization of an independent component analysis approach for artifact identification and removal in magnetoencephalographic signals. *Clinical Neurophysiology, 115*(5), 1220–1232.

Cao, J., Murata, N., Amari, S., Cichocki, A., & Takeda, T. (2003). A robust approach to independent component analysis of signals with high-level noise measurements. *IEEE Transactions on Neural Networks, 14*(3), 631–645.

Cichocki, A., & Amari, S. (2002). *Adaptive blind signal and image processing: Learning algorithms and applications*. West Sussex, England: Wiley.

Cichocki, A., Shishkin, S. L., Musha, T., Leonowicz, Z., Asada, T., & Kurachi, T. (2005) EEG filtering based on blind source separation (BSS) for early detection of Alzheimer's disease. *Clinical Neurophysiology, 116*(3), 729–737.

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing, 36*(3), 287–314.

Croft, R. J., & Barry, R. J. (2000). Removal of ocular artifact from the EEG: A review. *Neurophysiologie Clinique, 30*(1), 5–19.

Cruces-Alvarez, S., Cichocki, A., & Castedo-Ribas, L. (2000). An iterative inversion approach to blind source separation. *IEEE Transactions on Neural Networks, 11*(6), 1423–1437.

Delorme, A., Makeig, S., & Sejnowski, T. (2001). Automatic artifact rejection for EEG data using high-order statistics and independent component analysis. In *Proceedings of the 3rd International Workshop on ICA* (pp. 457–462). San Diego.

Escudero, J., Hornero, R., Abásolo, D., Poza, J., Fernández, A., & López, M. (2006) Rejection of artifact sources in magnetoencephalogram background activity using independent component analysis. In *Proceedings of the 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 5282–5285). New York.

Flexer, A., Bauer, H., Pripfl, J., & Dorffner, G. (2005). Using ICA for removal of ocular artifacts in EEG recorded from blind subjects. *Neural Networks, 18*(7), 998–1005.

Hämäläinen, M., Hari, R., Ilmoniemi, R. J., Knuutila, J., & Lounasmaa, O. V. (1993). Magnetoencephalography – theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of Modern Physics, 65*(2), 413–497.

Ikeda, S., & Toyama, K. (2000). Independent component analysis for noisy data – MEG data analysis. *Neural Networks, 13*(10), 1063–1074.

James, C. J. & Hesse, C. W. (2005). Independent component analysis for biomedical signals. *Physiological Measurement, 26*(1), R15–R39.

Jin, S. H., Jeong, J., Jeong, D. G., Kim, D. J., & Kim, S. Y. (2002). Nonlinear dynamics of the EEG separated by independent component analysis after sound and light stimulation. *Biological Cybernetics, 86*(5), 395–401.

Jousmäki, V., & Hari, R. (1996). Cardiac artifacts in magnetoencephalogram. *Journal of Clinical Neurophysiology, 13*(2), 172–176.

Joyce, C. A., Gorodnitsky, I. F., & Kutas, M. (2004). Automatic removal of eye movement and blink artifacts from EEG data using blind component separation. *Psycophysiology, 41*(2), 313–325.

Jung, T. P., Makeig, S., Humphries, C., Lee, T. W., McKeown, M. J., Iragui, V., & Sejnowski, T. J. (2000). Removing electroencephalographic artifacts by blind source separation. *Psycophysiology, 37*(2), 173–178.

Jutten, C., & Herault, J. (1991). Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing, 24*(1), 1–10.

Li, Y., Ma, Z., Lu, W., & Li, Y. (2006). Automatic removal of the eye blink artifact from EEG using an ICA-based template matching approach. *Physiological Measurement, 27*(4), 425–436.

Reyment, R., & Jöreskog, K. G. (1993). *Applied factor analysis in the natural sciences*. Cambridge: Cambridge University Press.

Sadasivan, P. K., & Dutt, D. N. (1996). SVD based technique for noise reduction in electroencephalographic signals. *Signal Processing, 55*(2), 179–189.

Sander, T. H., Wübbeler, G., Lueschow, A., Curio, G., & Trahms, L. (2002). Cardiac artifact subspace identification and elimination in cognitive MEG data using time-delayed decorrelation. *IEEE Transactions on Biomedical Engineering, 49*(4), 345–354.

Shayegh, F., & Erfanian, A. (2006) Real-time ocular artifacts suppression from EEG signals using an unsupervised adaptive blind source separation. In *Proceedings of the 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 5269–5272). New York.

Ting, K. H., Fung, P. C. W., Chang, C. Q., & Chan, F. H. Y. (2006). Automatic correction of artifact from single-trial event-related potentials by blind source separation using second order statistics only. *Medical Engineering & Physics, 28*(8), 780–794.

Vigário, R. N. (1997). Extraction of ocular artifacts from EEG using independent component analysis. *Electroencephalography and Clinical Neurophysiology, 103*(3), 395–404.

Vigário, R., Jousmäki, V., Hämäläinen, M., Hari, R., & Oja, E. (1998). Independent component analysis for identification of artifacts in magnetoencephalographic recordings. *Advances in Neural Information Processing Systems, 10*, 229–235. Cambridge, MA: MIT Press.

Vigário, R., & Oja, E. (2000). Independence: A new criterion for the analysis of the electromagnetic fields in the global brain? *Neural Networks, 13*(8), 891–907.

KEY TERMS

Artifact: Signal not produced by the brain, which appears in EEG or MEG recordings, and could bias the analyses of the neural activity. Some examples are the cardiac, ocular, muscular, and power line artifacts.

Blind Source Separation (BSS): The estimation of a set of unknown signals which have been mixed to produce a set of observed signals, with very little information about the former. Some common methods of BSS are ICA and PCA.

Electroencephalogram (EEG): Neurophysiologic recording of the voltage differences between different parts of the brain.

Higher-Order Statistics: Central moments of a statistical distribution with order higher than two.

Independent Component (IC): Each of the inner components into which ICA decomposes a multivariate signal.

Independent Component Analysis (ICA): A linear transformation for separating a multivariate signal into inner components, which are supposed to be as statistically independent as possible.

Magnetoencephalogram (MEG): Neurophysiologic recording of the magnetic fields produced by the currents of the pyramidal neurons that flow parallel to the skull.

Principal Component Analysis (PCA): A linear transformation for separating a multivariate signal into inner components in such a way that the greatest variance by any projection of the data lies on the first component, the second greatest variance on the second coordinate, and so on.

Applying Social Network Analysis in a Healthcare Setting

Salvatore Parise
Babson College, USA

INTRODUCTION

The people-to-people relationships where knowledge work actually gets performed in organizations are called social networks, and they may be in complete contradiction to the information flows expected, based on looking at the organizational chart of formal roles or titles. These informal or social networks are playing an increasingly important role in the healthcare industry, as medical and clinical knowledge needs to be shared effectively between people within and among healthcare organizations. Social network analysis (SNA) is a research methodology to analyze networks between people, groups, organizations, and systems within and across organizations (Wasserman & Faust, 1994). The results of the analysis inform the researcher of both the structure of the network, as well as the positions of nodes or people in the network. This article provides a description of how SNA can be applied in a healthcare setting.

Organizations are increasingly relying on networks of people, groups, and other institutions collaborating with each other to perform knowledge-based work. This is especially true in knowledge-intensive industries, such as business consulting, technology, research, and petroleum, where work is project-based in structure and sharing information within and between teams is essential to performance. Also, much of the critical work in businesses today requires tacit knowledge, which, by definition, is difficult to codify and resides in the key experts of the organization (Polanyi, 1983). Therefore, connections to these experts are required in order to leverage their expertise, past experiences, and institutional memory.

These informal or social networks are playing an increasingly important role in the healthcare industry. With the continued advancements in medical treatments, the rapid dissemination and sharing of this information becomes vital. While treatment descriptions can be codified and put into a database that physicians can access, often physicians will call on their personal relationships to understand the relevance, risks, and benefits of the latest medical knowledge. Physicians are turning to Web-based communities, such as sermo.com, to exchange, comment on, and rate others' postings on medical insights. Social networks are also vital to understanding the communication patterns, both between and within healthcare organizations. While traditionally operating as silos, these organizations recognize the need for improved access and sharing of clinical and medical research knowledge across their organizational boundaries. Within healthcare providers, social networks can reveal who are the key decision-makers or influencers pertaining to medical treatments, and the adoption of medical technology and information technology (IT), as well as determining if certain physicians are a bottleneck (or overburdened) when it comes to answering questions and disseminating information. Therefore a network perspective is vital to understanding how work gets performed in the healthcare industry.

BACKGROUND

There has been a dramatic rise recently in social network research in the management field (Borgatti & Foster, 2003). Network research has grown as a result of the importance of connections between people, groups, organizations, and IT systems. Network research has been used to study leadership (Brass & Krackhardt, 1999), entrepreneurship (Baron & Markman, 2003), knowledge management (Parise, Cross, & Davenport, 2006), individual performance (Mehra, Kilduff, & Brass, 2001), and team performance (Hansen, 1999). Social network analysis (SNA) is a structured methodology to analyze networks within and across organizations. The results of the analysis inform the researcher of both the structure of the network as well as people's positions within the network. Based on these findings, organizations can then develop interventions to produce the desired network effects. There has been limited research in the healthcare setting, using SNA as a methodology. SNA has been used to study the interaction patterns in primary care practices (Scott, Tallia, Crosson, Orzano, Stroebel, DiCicco-Bloom, O'Malley, Shaw, & Crabtree, 2005), identify influential individuals who are critical to the successful implementation of medical informatics applications (Anderson, 2005), and study the relationship between communication density and the use of an electronic medical record system by nurse practitioners and physicians' assistants (Tallia, Stange, McDaniel, Aita, Miller, & Crabtree, 2003).

SNA allows us to analyze the relationships among actors or nodes in a network. Nodes can be people, organizations, and IT systems. Connections between nodes are called links and determine the relationships between nodes. Two common representations of networks are bounded networks and ego networks (Wasserman & Faust, 1994; Scott, 2000). In a bounded network, the nodes in the network are predetermined. When using bounded networks, the researcher will often look for silos or fragments in the network. This occurs when there is a breakdown in communication between structures, such as between departments or divisions. Attribute data, such as current job tenure, company tenure, job level, location, and functional area are often collected for each node in the network. This allows the researcher to analyze the structure of the bounded network, and determine if any of these dimensions are having a significant impact on its structure. Ego networks are those generated based on the perspective of an individual node or person in the network. Typically, in an ego network analysis, each respondent is asked to list nodes or actors pertaining to them specifically. The SNA researcher can then derive the network, based on the nodes listed by all respondents, as well as study any biases or gaps in an individual's personal network.

SNA researchers will typically analyze both the structure of the network as well as the role of individual nodes within the network. Density and centralization are two common metrics used to describe the entire network. Density is defined as the ratio of observed links or connections in a network to the maximum number of links possible if every node in the network was connected (Wasserman & Faust, 1994). Centralization refers to the degree to which links are concentrated in the network. The centralization metric is very useful in analyzing the degree to which there are "dominant" nodes in the network.

In addition to analyzing the network structure, the other major insight from doing an SNA analysis is to understand a node's role. Certain nodes in the network may be more central resulting in a more advantageous position. Individuals who are more central in the network have been found to have greater influence (Brass, 1984; Brass & Burkhardt, 1992), and have greater access and control of relevant information (Brass & Burkhardt, 1993; Krackhardt, 1990; Umphress, Labianca, Brass, Kass, & Scholten, 2003). In terms of identifying nodes with power or influence, central degree, and central betweeness metrics can be used. Degree centrality refers to the number of links going into ("in-degree") or coming out of ("out-degree") a node in a network (Freeman, 1979). Brokers in the context of SNA are used to describe nodes in the network that connect different subgroups. They may not have the most direct links, but they act as a conduit between departments, locations, and hierarchies. One common way to identify brokers is the betweeness centrality metric. Betweeness centrality for a particular node refers to the number of links that each other node in the network needs to go through that particular node to reach any other node in the network (Freeman, 1979).

SNA USE IN HEALTHCARE

Researchers can apply both a network perspective and the SNA methods to better understand the healthcare setting. In particular, there are two areas in healthcare that can benefit from a network approach: (1) analyzing medical and clinical information flows within and between healthcare organizations, and (2) identifying key "influencers" with respect to adopting medical treatments, medical technologies, and IT, or recommending referrals.

SNA can be used as an analytical tool to study interaction patterns within medical practices. It is critical to understand who the medical staff turn to for help regarding medical treatments, unstructured and complex problems, or decision-making. Great variations in organizational design exist in primary care practices (Tallia et al., 2003). Scott et al. (2005), for example, used SNA to study the decision-making patterns in two different medical practices, and found two vastly different organizational structures: one practice had a hierarchical structure where the practice leader had final decision-rights, while the other practice had a much more collaborative decision-making process where staff had the authority to make decisions on their own. Also, physician-to-physician communication is critical to the success of outpatient referral (Gandhi, Sittig, Franklin, Sussman, Fairchild, & Bates, 2000). This process requires the transfer of relevant clinical information in both directions between the physician and specialist. Furthermore, primary care physicians are often regarded as an obstacle or gatekeeper with respect to referring patients to specialist care (Scherger, 2002). An SNA could determine if any bottlenecks exist between the primary care physician and the specialist in the referral process.

SNA can be very effective at identifying influential individuals in the implementation of medical informatics applications and in the patient referral process. Suppliers of medical equipment face new challenges as new players are having a more active role in the assessment and purchase decision process (IBM E-Health Transformation report, 2002). Also, in terms of assessing the effectiveness of the referral process, one would need to know if there are only a few physicians making most of the referral decisions. For example, Anderson (2005) found that a group of physicians controlled the referral process, and were then enlisted in the planning and implementation of a new IT system.

CASE STUDY: SNA AT A MEDICAL CENTER

An SNA study was conducted at a medical center to analyze the collaboration between primary care physicians and specialist physicians. The center provided comprehensive adult and adolescent primary care. It consisted of several departments and practice areas, including primary care and major specialty areas, such as cardiology and gastroenterology. Most of the staff resided in two buildings located a few hundred feet apart in a campus setting. The research sample consisted of two practice areas: the primary care practice and a large specialist practice. A total of 57 respondents took the survey, including primary care physicians (PP), specialist physicians (SP), office managers (OM), administrative staff (ADM), nurses (N), physician assistants (PA), and lab personnel (LAB). All survey network data was analyzed using UCINET VI (Borgatti, Everett, & Freeman, 2002).

Three bounded networks were used that focused on information sharing, advice regarding complex medical problems, and the referral process. The questions used in the survey include:

- **Information:** *Please indicate the extent to which the people listed below provide you with information you use to accomplish your work.* A “0” response indicated the respondent did not know the person. A “1” (very infrequent) to “6” (very frequent) likert scale was used.
- **Advice:** *Please indicate the extent to which the people listed below provide you with advice regarding unstructured, complex medical problems.* A “1” (very infrequent) to “6” (very frequent) likert scale was used.
- **Referrals:** *Please indicate the extent to which the people listed below provide you with patient referrals.* A “1” (very infrequent) to “6” (very frequent) likert scale was used.

The information, advice, and referral networks were dichotomized at responses greater than “4,” so as to only use strong responses as links in the networks. All 57 respondents answered the Information network question, and only the 35 physicians were given the Advice and Referral network questions.

Ego networks were also examined to get a sense of where physicians obtained critical knowledge and what expertise areas were critical to perform their job. Each physician respondent was asked to:

- *Identify (up to 10) critical sources of information that help you with your job.* These sources can be individuals (inside or outside the medical center) as well as IT systems (such as on-line medical communities).
- *Identify the major (up to 10) expertise areas that are critical for you to perform your job.* For each expertise area, the respondent was asked to associate any of the sources of information to the expertise area.

Table 1 shows the overall network densities, as well as the dominant employees in the information and advice networks. The information network had an overall density of 12.7%. The top central connectors were PP4 and PP8, as 23 people in the network sought out PP4 for information, while 20 respondents relied on

Table 1. Key influencers () indicate in-degree scores

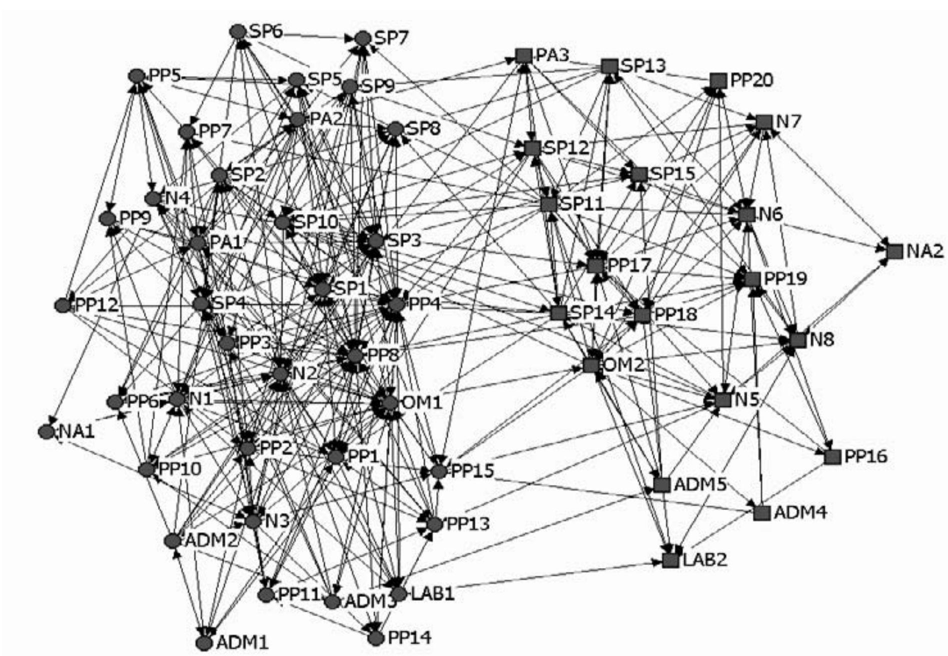
Network	Density	Central Connectors	Brokers
Information	12.7%	PP4 (23) PP8 (20) SP3 (16) OM1 (15)	OM1 OM2 PP4 SP10
Advice	8.5%	SP3 (17) SP12 (13) PP8 (11)	SP12 PP8 SP14 SP3 PP18

PP8 for information. The top brokers in the information network (as measured by betweeness scores) were the two office managers (OM1 and OM2) and PP4. The most surprising find was the fragmentation caused by building location. Figure 1 depicts the information network, and there is a clear divide in information sharing among employees between the two buildings. Of the 670 total links, 592 were within buildings, and only 78 links existed across the two buildings.

Figure 2 depicts the advice network in solving complex medical problems. The arrows in the network point to the source of the advice. The advice network was much less dense (8.5%) than the information network, and was much more concentrated with 43% centralization, compared with 28% for the information network. In other words, many of the respondents relied on three physicians to help solve complex problems: two specialists (SP3 and SP12), and one primary care physician (PP8).

Figure 3 depicts the referral network. The arrows in the network point to the source of the referrals. The source of the referrals was dominated by three primary care physicians: PP4, PP8, and PP17. These three physicians knew most of the specialists and their specific

Figure 1. Information network (the arrows in the network point to the source of work-related information)



Building 1

Building 2

Figure 2. Advice network (the arrows in the network point to the source of advice regarding complex medical problems)

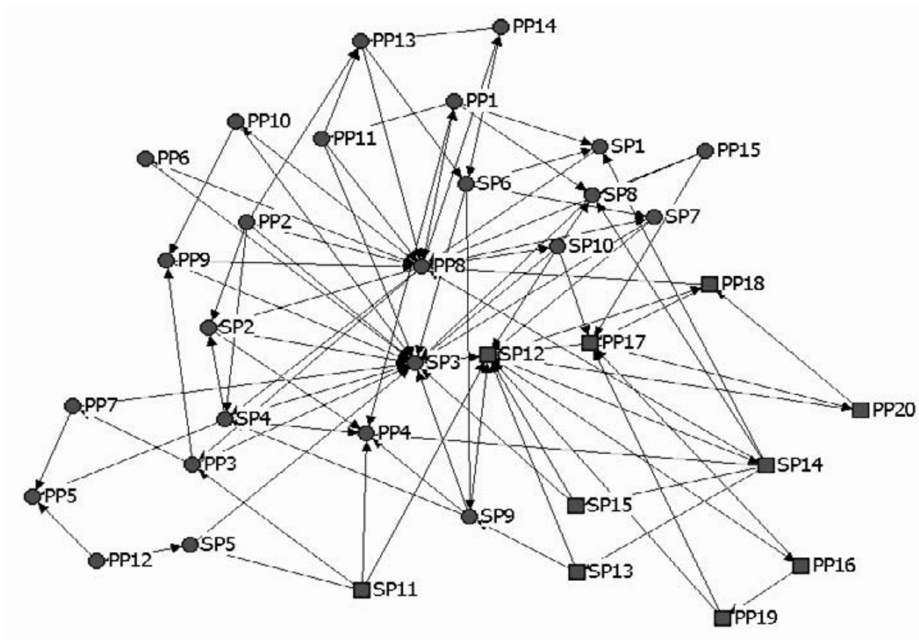
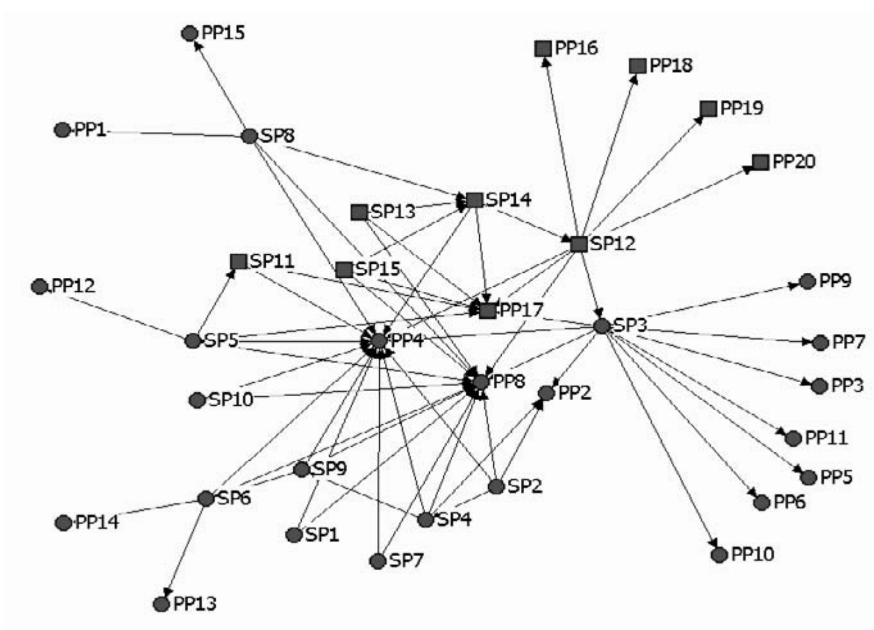


Figure 3. Referral network (the arrows in the network point to the source of the referrals)



areas of expertise. Specialists indicated all three were effective in both deciding when a referral was needed, as well as providing effective content in the referral letter itself. Other primary care physicians often relied on PP4, PP8, and PP17 for advice regarding referrals. Follow-up interviews clearly revealed other physicians were not as effective in the referral process and most referred patients to only a single specialist with whom they were somewhat familiar and comfortable communicating, but was not necessarily the ideal specialist, based on the patient's condition.

Finally, the ego network yielded interesting findings, especially when comparing the personal network of the well-connected physicians with those that were on the periphery. Several of the well-connected physicians had listed many diverse sources of information, and for each of their expertise needs, they had several knowledge sources from which to choose (including physicians from other hospitals). Meanwhile, many of the peripheral physicians had personal networks that resembled primary care physician PP12's personal network. PP12 had been with the medical center for just about one year. She listed the following expertise areas as critical to her job:

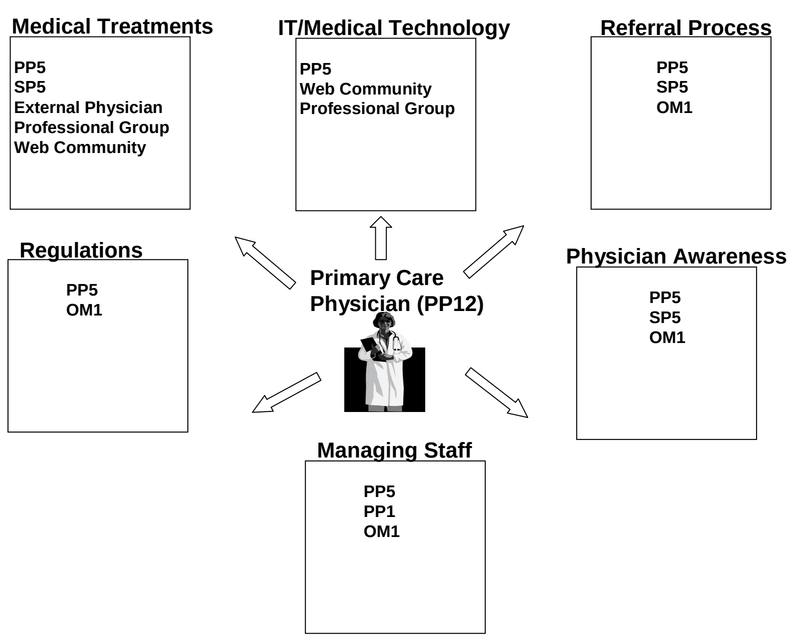
- Latest medical treatment knowledge
- IT systems and medical technology knowledge
- Referral process understanding
- Regulations/legal issues knowledge
- Awareness of other physicians' expertise areas
- Managing staff

As seen in Figure 4, her personal network was very limited and insular, as she relied too heavily on PP5, SP5, and OM1 to get her work done. She indicated in interviews that she was frustrated with doing too much "busy work," and did not feel well-connected with other physicians at the center. She admitted she was unaware of the specific areas of expertise of many of the physicians, so as a result, she relied heavily on only a few people at the center.

From these SNA findings, the office managers developed some interventions to improve collaboration and information sharing among the staff:

- It was clear from the SNA study that PP4, PP8, SP3, and SP12 were very influential at the medical center, since each of them held very central positions in the information, advice, and referral networks. All four of these physicians had over 10 years of tenure at the center. They all had built relationships with doctors throughout the industry, and they were well represented at professional conferences, and were members of professional

Figure 4. Physician PP12's personal network



groups and communities. However, they indicated that one source of frustration was they thought too many physicians and nurses came to them for routine questions or medical information, which could easily be handled by other staff members. Therefore, one action item was to free-up time for these physicians by delegating decision-rights for certain procedures and tasks to other physicians and nurses. In this way, they could focus on the complex medical problems.

- Another way to leverage these four influencers (PP4, PP8, SP3, and SP12) was to help them develop staff, especially new doctors coming into the center. Periodic meetings (e.g., once a month) could be scheduled between a newcomer and an appointed influencer physician. The meetings could be used as a forum for the newcomer physicians to: ask any questions about medical treatments and procedures; help build their awareness of the other doctors and their expertise areas; better understand the referral process, specifically which specialists should be contacted for what procedures and how to best communicate with certain specialists; help build their personal network by having experienced physicians introduce them to their own personal contacts and professional groups/communities; and create a trusting and sharing community with other doctors at the center.
- The office managers also realized they needed to leverage the influencer physicians in terms of planning and implementing their IT plans for the center. The goal was to have an interactive Web site and intranet for the center. The center staff would use the intranet to send emails, to do prescription refills, to schedule patient appointments, to access medical databases and communities, and to create expertise profiles for each of the staff members. The hope was that physicians could also use the intranet to develop their own content and to share their own experiences regarding the latest medical technologies and treatments. The office managers also wanted to make the center a "paperless" office by digitizing clinical records and making all patient information available either through the intranet or handheld computers. But, to achieve this vision, the office managers now realized they needed to get the influencer physicians to be a strong proponent of their IT plans, so they were invited to be part of this IT steering committee. This would also be a good opportunity to get some of the new doctors (who were more familiar with using IT) to be part of the IT committee.
- The referral network clearly showed the need for better communication between the primary care physicians and specialists, as that network was dominated by only three of the 20 primary care physicians. Part of the problem stemmed from a lack of awareness of what each of the specialists did, and the other part of the challenge was an understanding of what clinical information was expected to be transferred between the primary care and specialist physician. The hope was that IT could aid the referral process. Having an expertise profile for each of the physicians on the intranet could help build awareness. Specialists could also post information about their expertise areas, including information regarding a disease, its treatment, and caring for the patient. IT could also help automate the referral process by generating referral letters that could be sent via email.
- Finally, one of the most surprising findings was how much fragmentation there existed in the information network as a result of the staff's building location. One idea was to rotate work assignments across the two buildings, especially for the primary care physicians and their staff. For example, primary care physicians (especially newcomers) could work certain days of the week in each of the two buildings. The physical proximity would help the physicians build awareness and relationships with other doctors, including specialists.

FUTURE TRENDS

SNA continues to be a popular research methodology in understanding information flows among people in knowledge-intensive organizations. Going forward, there are many opportunities to incorporate a network perspective, and to leverage SNA in the healthcare setting. One specific area is to understand the information flows between healthcare organizations, such as hospitals, medical centers, pharmaceutical companies, and insurers. As the healthcare industry continues to evolve and change to be more efficient and effective in delivering patient care, both medical and clinical information need to be more transparent and accessible across organizations. SNA could be an effective means to map collaboration and information flows at an industry-level.

Another area that could benefit from incorporating a network perspective is to understand how physicians are using online medical communities to gain access to medical knowledge. Physicians today are able to tap into a larger network of physicians through blogs, wikis, and discussion threads on specific medical topics. These communities can exist both within and outside organizational affiliations. An ego network could be used to analyze a physician's personal network in terms of using electronic medical communities as a source of their medical knowledge. In this way, popular or effective communities could be shared with other physicians. Or, a network map could be created that identifies physicians as nodes, and comments or responses to physicians in these communities as links between physicians. In this way, central physicians could be identified based on their frequency of postings, and to which physicians they are responding to. The effectiveness of the postings (based on a user rating system) could also be incorporated as part of the network map.

CONCLUSION

This article has provided a description on how a network perspective, and in particular, a social network analysis (SNA) methodology could be applied in a healthcare setting. As illustrated in the case study, SNA allows the researcher to identify areas of fragmentation in the network as well as central people in the network. Based on this analysis, interventions can then be developed to improve the network structure and network roles, ultimately resulting in more effective collaboration. Going forward, incorporating a network perspective to understand the healthcare setting is important as collaboration among different stakeholders and sharing of medical and clinical knowledge remains a critical strategy.

REFERENCES

Anderson, J.G. (2005). Evaluation in health informatics: Social network analysis. In J.G. Anderson, & C.E. Aydin (Eds.), *Evaluating the Organizational Impact of Healthcare Information Systems* (pp. 189–204). London: Springer.

Baron, R.A., & Markman, G.D. (2003). Beyond social capital: The role of entrepreneurs' social competence in their financial success. *Journal of Business Venturing*, *18*(1), 41–60.

Borgatti, S.P., Everett, M., & Freeman, L. (2002). *UCINET 6 for Windows*. Natick, MA: Analytic Technologies.

Borgatti, S.P., & Foster, P.C. (2003). The network paradigm in organizational research: A review and typology. *Journal of Management*, *29*(6), 991–1013.

Brass, D.J. (1984). Being in the right place: A structural analysis of individual influence in an organization. *Administrative Science Quarterly*, *29*, 518–539.

Brass, D.J., & Burkhardt, M.E. (1992). Centrality and power in organizations. In N. Nohria, & R.G. Eccles (Eds.), *Networks and organizations: Structure, form and action* (pp. 191–215). Cambridge: Harvard Business School Press.

Brass, D.J., & Burkhardt, M.E. (1993). Potential power and power use: An investigation of structure and behavior. *Academy of Management Journal*, *36*, 441–470.

Brass D.J., & Krackhardt, D. (1999). The social capital of 21st century leaders. In J.G. Hunt, & R.L. Phillips (Eds.), *Out-of-the-box leadership* (pp. 179–194). Stamford, CT: JAI Press.

E-Health Transformation. (2002, October). *Managing healthcare in a networked world*. IBM Business Consulting Services report, G510-9212-00.

Freeman, L.C. (1979). Centrality in social networks: Conceptual clarification. *Social Networks*, *1*, 215–239.

Gandhi, T.K., Sittig, D.F., Franklin, M., Sussman, A.J., Fairchild, D.G., & Bates, D.W. (2000). Communication breakdown in the outpatient referral process. *Journal of General Internal Medicine*, *15*, 626–631.

Hansen, M.T. (1999). The search-transfer problem: The role of weak ties in sharing knowledge across organization subunits. *Administrative Science Quarterly*, *44*(1), 82–111.

Krackhardt, D. (1990). Assessing the political landscape: Structure, cognition, and power in organizations. *Administrative Science Quarterly, 35,* 342–369.

Mehra, A., Kilduff, M., & Brass, D.J. (2001). The social networks of high and low self-monitors: Implications for workplace performance. *Administrative Science Quarterly*, *46*(1), 121–146.

Parise, S., Cross, R., & Davenport, T.H. (2006). Strategies for preventing a knowledge-loss crisis. *MIT Sloan Management Review*, *47*(4), 31–38.

Polanyi, M. (1983). *The tacit dimension*. New York: Doubleday.

Scherger, J.E. (2002). Challenges and opportunities for primary care in 2002. *Medscape Family Medicine/Primary Care*, *4*(1).

Scott, J. (2000). *Social network analysis: A handbook*. London: Sage.

Scott, J., Tallia, A., Crosson, J.C., Orzano, A.J., Stroebel, C., DiCicco-Bloom, B., O'Malley, D., Shaw, E., & Crabtree, B. (2005). Social network analysis as an analytical tool for interaction patterns in primary care. *Annals of Family Medicine*, *3*(5), 443–448.

Tallia, A.F., Stange, K.C., McDaniel, R.R., Aita, V.A., Miller, W.L., & Crabtree, B.F. (2003). Understanding organizational designs of primary care practices. *Journal of Healthcare Management*, *48*(1), 45–61.

Umphress, E.E., Labianca, G., Brass, D.J., Kass, E., & Scholten, L. (2003). The roles of instrumental and expressive social ties in employees' perceptions of organizational justice. *Organization Science*, *14*, 738–753.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Boston: Cambridge University Press.

KEY TERMS

Broker: In the context of SNA, brokers are used to describe nodes in the network that connect different subgroups.

Degree Centrality: Refers to the number of links going into ("in-degree") or coming out of ("out-degree") a node in a network.

Information Network: Information networks map who people turn to for information regarding work-related activities.

Network Fragmentation: In the context of SNA, network fragmentation occurs where silos exist in the network and information is not being transferred across functional, hierarchical, cultural, or geographic boundaries.

Social Network: The people-to-people relationships in organizations where knowledge work actually gets performed. The social network may be in complete contradiction to the information flows expected, based on the organizational chart of formal roles or titles.

Social Network Analysis: A research methodology to analyze networks between people, groups, organizations, and systems within and across organizations.

Tacit Knowledge: Knowledge that is difficult to codify, and often resides in the key experts of the organization.

Applying the Quality Loss Function in Healthcare

Elizabeth Cudney
Missouri University of Science and Technology, USA

Bonnie Paris
University of Wisconsin-Madison, USA

INTRODUCTION

Using the quadratic loss function is one way to quantify a fundamental value in the provision of health care service: we must provide the best care and best service to every patient, every time. Sole reliance on specification limits leads to a focus on "acceptable" performance rather than "ideal" performance. This chapter presents the application of the quadratic loss function to quantify improvement opportunities in the health care industry.

The quadratic loss function (QLF), also known as the quality loss function, is a metric developed by Genichi Taguchi, which focuses on achieving the target value rather than focusing on performance within the wider specification limits. Using the quadratic loss function allows the Six Sigma team to quantify improvement opportunities in monetary terms, the language of upper management. The quadratic loss function translates variability into economic terms by calculating the relationship between performance and financial outcome. The general quadratic loss function is shown in Equation 1.

(1)

*Loss at any point (L) = (monetary constant) * (average – target)*^{2}

The quadratic loss function is used to determine the average loss per product or encounter, and it enables Six Sigma teams to focus on performance relative to target and avoid the goal-post mentality. The loss function approximates the long-term loss from performance failures and encourages continuous improvement. The quadratic loss function is helpful both as a philosophical approach and as a quantitative method. Figure 1 illustrates the quadratic loss function.

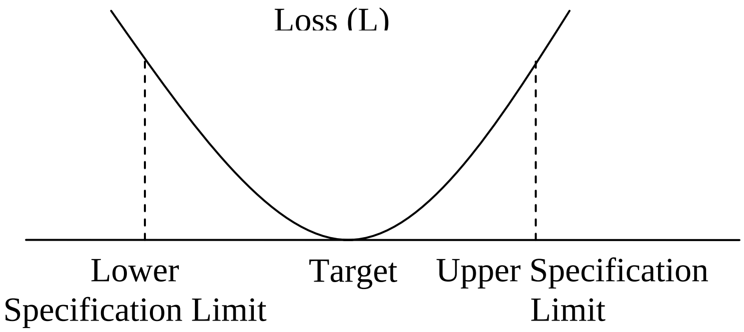
BACKGROUND

Many attempts have been made to make a business case for quality by examining cost, but a thorough review of literature shows no evidence of a sustained cost of quality system implementation in health care. Several researchers have commented on the paucity of information on this topic (Palladoro, 1997; Bozanich White, 1999), and in 2003, a collaborative, including Donald Berwick, M.D., president and CEO of the Institute for Healthcare Improvement, "sought to erase the notion that quality doesn't pay by setting up five case studies that would establish a 'business case for quality.'" All five failed to do so (Robeznieks, 2003).

One reason for this failure may be the fragmented information sources from which health care quality data are derived. Campanella (1990) cautions that "a real danger lies in finding and collecting only a small portion of the costs involved and having it represented as the total." This has likely been the case with previous attempts to cost-justify preventive measures. Getting the resources to collect the right data requires leadership commitment and technical expertise, especially in situations where data must be harvested from several sources and brought together for analysis. When performing a cost of quality analysis, it is preferable to use an activity-based cost (ABC) approach to gathering data because the ABC costs are analyzed in terms of work activities (Cokins, 1997), but hospital accounting is a tangled web of allocations developed to support cost reimbursement and was not designed to assist managers in decision-making (Bozanich White, 1999).

Many models consider the costs and benefits of providing care that is given, but do not address the cost associated with failure to provide care. There is a great deal of work remaining in the comprehensive application and adoption of cost of quality techniques to health care, and there is some indication that the current state

Figure 1. Quadratic loss function



of quality cost thinking in health care is similar to that of industry 60 years ago. The traditional view holds that high quality care costs more because it was seen as doing everything possible for the patient, whether or not the additional tests or extended hospital stays improved the clinical outcome (Burda, 1992).

In order to discuss the cost of quality in measurable terms, one must first have agreement on what is meant by quality. Safety is another aspect of quality, where patient safety is the avoidance of accidental harm to the patient (Kohn, Corrigan, & Donaldson, 2000). Some define safety more broadly to include avoidance of all harm to the patient. David Aquilina (1992) developed a framework for understanding health care quality that looks at health care quality as multidimensional, overlapping concepts. Industry usually defines quality in terms of customer requirements, but this can prove difficult in health care due to the complexity of the customer relationship and the complexity of the service (Aquilina, 1992).

The relationship between health care quality, however defined, is not easily related to the health care industry's bottom line. As in other industries, cost, revenue, and quality are codependent. However, in the current health care environment, even the link between throughput and revenue is indirect. There is a disconnect between the provision of service and payment of service due to a myriad of factors, such as government intervention and ability to pay vs. professional obligation to provide care.

The loss function makes the point that if performance is off-target in any way, there is a loss to society as a result. The loss may be experienced by the provider, the consumer, the environment, and so forth, but there is a cost associated with imperfection. At the point where total cost is minimized, quality is highest. In other words, the highest quality care is also the care with the lowest overall cost, when all costs are accounted for.

Although there are different opinions as to how quality and cost of quality relate to achieve an optimal level, there is agreement that money spent on prevention yields the highest return compared to any other quality cost component. This may be especially true in health care, where the prevention of an error may be possible, but once made, there is no opportunity to correct the error. An example of this is wrong site surgery, where a surgical procedure may be performed on the wrong limb. As a corrective action, the "right" surgery may be done, but the "wrong" surgery cannot be undone. Unlike other service industries, this "extra" service can cause permanent, irreversible harm to the patient (customer).

Empirical measurement of the care provided in terms of improved patient outcomes on a case-by-case basis is a labor-intensive but accepted method of evaluating the effectiveness of care. Controlling for measures such as type of illness, severity of illness, and other clinical factors on a case-by-case basis and evaluating the cost/risk and benefit of providing the care in each circumstance allows clinicians to develop evidence-based practice patterns. Even when evidence-based medicine is used and clinical outcomes are measured, these pieces of information do not add up to a picture of systemwide performance.

Evidence-based medicine is one standard; "care maps" for certain patient populations developed by interdisciplinary teams, specifications by government, health plans, and the like are growing in number. Standards in health care are not static, and the bar is always rising as medical knowledge increases. However, it is not always inappropriate to violate one standard in favor of another, so long as the standard is violated in the best interest of the patient (Laffel & Blumenthal, 1992).

An "ideal" cost of quality for health care could be estimated based on the principle that where total cost to society as a whole is the least, quality is the highest. Access to care, or the ability to obtain health care services, may have a significant impact on the "hidden" costs of "poor" quality health care. Likewise, individual participation in taking preventative measures, such as healthy diet and exercise, may also contribute significantly to overall demand and cost of health care. From a societal standpoint, these are individuals who may have prevented the onset of or complications of disease with earlier intervention but have instead been "reworked" through the health care system.

QUALITY LOSS FUNCTION

Genichi Taguchi, a Japanese engineer, developed the quality loss function based on the economic consequences of not meeting target specifications. Taguchi defined quality as the "[avoidance of] loss a product causes to society after being shipped, other than any losses caused by its intrinsic functions" (Evans, 1999). Losses to society encompass costs incurred by the manufacturer, the customer, and anyone affected by product deficiencies.

The quality loss function unifies quality and cost to drive on-target engineering. It also relates economic and engineering terms in one model. The function enables optimization of costs through the production and use of a product (Fowlkes & Creveling, 1995).

The loss to society includes costs incurred due to a product or service not meeting customer expectations, not meeting performance characteristics, and having harmful side-effects. The quality loss function associates a dollar value to the current state of quality for a product or process. The quality loss function approximates losses due to scrap, rework, poor performance, lack of customer satisfaction, and so forth. This dollar amount can be used to identify areas for improvement and evaluate improvement efforts.

Taguchi relates the monetary loss to quantifiable product or service characteristics. This translates the language of an engineer into management terms. The quality loss function focuses on target values rather than specifications for process output. When nominal specifications are achieved, the products are consistent, and there is no loss to society (Breyfogle, 2003). Any deviation from target incurs a loss to society. The loss increases as the measured responses move further away from the target value. The measured loss L(y) for a single product is estimated as shown in Equation 2.

$$L(y) = k(y - T)^2 \qquad (2)$$

where,

- L(y) is the loss in dollars,
- y is the measured response,
- T is the target value of the product's response, and
- k is the quality loss coefficient.

The quality loss coefficient, k, is determined using customer tolerance and the economic consequence. The economic consequence, A_0, is the cost to replace or repair the product. The associated costs include losses incurred by the manufacturer, customer, or third party. The customer tolerance is the point at which the product reaches unacceptable performance.

$$k = \frac{A_0}{\Delta_0^2} \qquad (3)$$

where,

- A_0 is the economic consequence of failure, and
- Δ_0 is the functional limits or customer tolerance for the measured response.

The quality loss function can also be used to determine the average loss per product. The expected loss, E(L), is used to depict the average loss. To reduce the expected loss, the variability and deviation from target must be reduced (Kiemele, 1999). Equation 4 shows the expected loss.

$$E(L) = k\left(\sigma_y^2 + (\bar{y} - T)^2\right) \qquad (4)$$

where,

- σ_y^2 is the variation,
- $\bar{y}$ is the response average,
- T is the target value, and
- k is the quality loss coefficient.

FUTURE TRENDS

Six Sigma is a customer-focused continuous improvement strategy and discipline that minimizes defects and variation toward an achievement of 3.4 defects per million opportunities. However, this target of 3.4 defects per million may create a "goalpost mentality" that everything within specification limits is equally good. In this traditional thinking, quality is achieved through conforming to specifications. However, the reality is that products just within the specification limits are not just as good as products that are directly on target. The quality loss function also addresses the concern that a defect rate of 3.4 per million opportunities may not be an appropriate goal for all processes, such as compliance with the medical library book check-out procedure.

Specifications have no value for the customer. The customer expects performance and does not care about specifications. Specifications use the old quality mentality to create "goalposts" for product acceptance for the supplier. Specification limits create a natural tendency to create products inside the limits and assume that all products inside the limits are equal. Specifications do not address the associated loss for not meeting the target.

In a hypercompetitive market with highly educated and discerning customers, companies cannot afford a uniform distribution of products. By only focusing on zero defects, companies will also go out of business because they are not using metrics that correlate to customer satisfaction.

These concepts have been used extensively in manufacturing, but they also apply to health care processes. For example, a focus study finds that patients are willing to wait 15 minutes to be seen by the physician for routine appointments. One approach is to look at the percentage of wait times of 15 minutes or less and consider these experiences to meet the customer expectation. However, a wait time of 14.9 minutes is considered as acceptable as a wait time of 2 minutes with this traditional approach.

Consider a hypothetical example of a surgical procedure with a typical hospital stay of two to five days with an average length of stay of 2.5 days. The target length of stay is 3 days because patients staying less than 3 days have an increased readmission rate, and patients staying longer than 3 days had complications and a prolonged recovery period.

A simulated data set containing 30 observations was developed to illustrate the differences between the traditional approach and the quality loss function approach. Using the traditional approach, the lower specification limit and upper specification limit would be two and five days, and 20% of the observations were considered outside of the target with 10% below the lower specification limit and 10% above the specification limit. This is shown in the Figure 3.

However, using the quality loss function approach, it is determined that only 16.7% of the observations are on target with a three-day length of stay. Using the quality loss approach, we are not meeting target more than 80% of the time.

Traditional thinking looks at stays less than three days as a cost savings, but due to increased patient

Figure 2. Perception of quality through meeting specifications

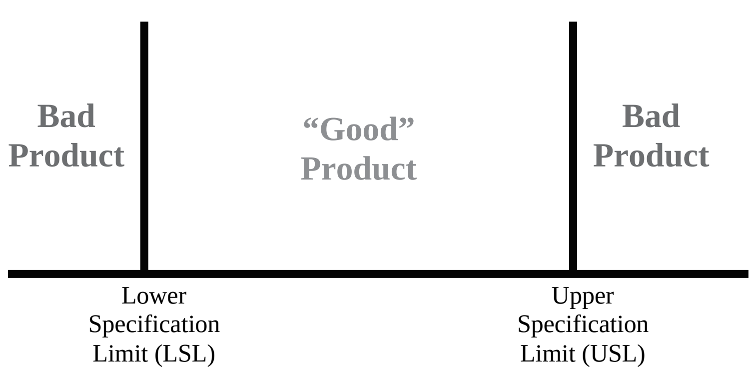
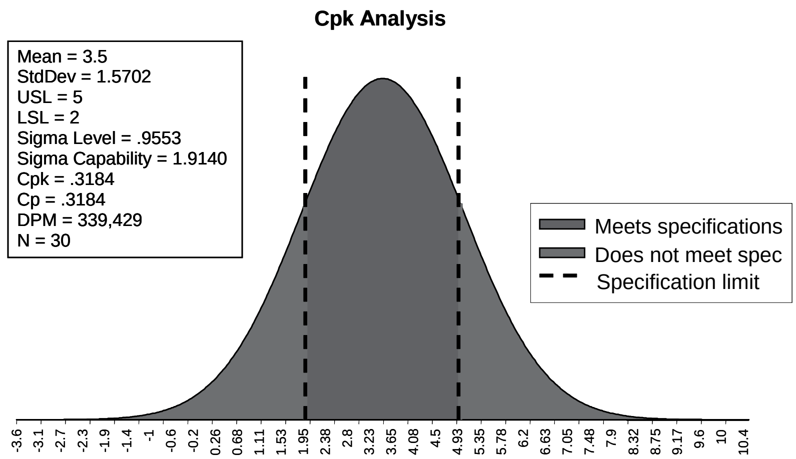


Figure 3. Traditional approach using CpK



readmissions, this is not the case. There is a cost associated with early release, and this becomes apparent when we expand our focus from the patient encounter to the episode of care and case management of the patient's health, whether in or out of the hospital.

Suppose we calculate the monetary constant for the quadratic loss function as \$212.50, based on the average loss per patient and the patient's tolerance for a length of stay that is too short or too long. The average length of stay is known to be 2.5 days with a standard deviation of 1.57 days, and the target is three days. The average loss for the sample set can be found using Equation 4.

$$E(L) = k\left(\sigma_y^2 + (\bar{y} - T)^2\right)$$

E(L) = \$212.50 [(1.57*days*)2 + (2.5*days* - 3*days*)2] = \$576.92

We can calculate the total loss by multiplying the average loss by the case volume. If there are 350 procedures per year, then the total loss is \$201,922 per year.

Currently, use of the quality loss function is restricted to situations where quantitative information about product or service performance is available. To use the quality loss functions in situations where only qualitative data are available, a numeric measure of effectiveness or ranking can be used.

CONCLUSION

Quality is based on achieving a target value set by the customer, and any time there is deviation from the target, there is a loss. This loss can take many forms, including customer dissatisfaction, poor or unreliable performance, and rework, which are equated with monetary loss. Using the quality loss function can assist in quantifying and understanding this loss, in support of continuous improvement.

There is a great deal of work remaining in the comprehensive application and adoption of the quality loss function health care. However, the work will provide valuable information to decision makers and therefore assist in improving the cost and quality of care provided.

REFERENCES

Aquilina, D. (1992). Alternative approaches to defining quality in medical care. *Quality Assurance in Health Care Services*, 306–312.

Bozanich White, J.B. (1999). *Identification of quality management activities and their associated costs and effectiveness in acute care hospitals.* Doctoral dissertation, George Mason University.

Breyfogle, F. (2003). *Implementing six sigma*. NJ: John Wiley & Sons, Inc.

Burda, D. (1992). Providers look to industry for quality models. Qu*ality Assurance in Health Care Services*, 7–14.

Campanella, J. (Ed.). (1990). *Principles of quality costs: Principles, implementation, and use*. Milwaukee: ASQC Quality Press.

Cokins, G. (1997). If activity based costing is the answer, what is the question? *IIE Solutions*, 39–42.

Evans, J., & Lindsay, W. (1999). *The management and control of quality*. OH: South-Western College Publishing.

Fowlkes, W., & Creveling, C. (1995). *Engineering methods for robust product design*. MA: Addison-Wesley.

Kiemele, M. (1999). Basic statistics. Colorado Springs, CO: Air Academy Press, LLC.

Kohn, L.T., Corrigan, J.M., & Donaldson, M.S. (2000). *To err is human: Building a safer health system*. Committee on Quality of Health Care in America, Institute of Medicine. Washington, DC: National Academy Press.

Laffel, G., & Blumenthal, D. (1992). The case for using industrial quality management science in health care organizations. *Quality Assurance in Health Care Services*, 36–46.

Palladoro, V.A. (1997). *An investigation of continuous quality improvement in healthcare: Is there an evaluation strategy?* Doctoral dissertation, University of Rochester.

Robeznieks, A. (2003). *Quality care not tied to financial return: The search continues for a "business case" for improving health care quality*. Retrieved from www.amednews.com.

A

KEY TERMS

Fragmented Information Sources: Separate sources of data and information that require significant effort and skill to combine in a meaningful way.

Goalpost Mentality: The thought that performance between targets is always and equally acceptable.

Health Care Quality: Attributes of output of the health care process, including clinical outcomes, access to care, satisfaction with care, and cost of care.

Hidden Costs: Nonobvious costs of providing or failing to provide care, including the value of lost patient and provider time.

Tolerance: An acceptable level of imperfection in service received as defined by the customer.

Appreciation Level and Organizational Performance

Murako Saito
Waseda University, Japan

INTRODUCTION

Corporate environment changing with the advent of information technology and with diversified organization has been inquired to redesign to transform into an intelligent and innovative organization so that corporate vision and organizational goals in the subsystem of corporate are provided for being shared by multidisciplinary workers and other stakeholders, and corporate social responsibility can also be shared by all the participants. System matching between developing levels in technological systems and also cognitive fit in the levels of individual, team/group, and organization are required not only to maintain good balance, but also to enhance operational and organizational resilience in making appropriate shifts to adapt to a changing social environment. To make the shift successfully, cognitive fit or value alignment between individuals and collectives in the environment of advanced technologies is to be studied, so that corporate social responsibility or corporate prosperity is shared by all the stakeholders.

Organizational performance and the performances in the level of heterogeneous team/group of an organizational subunit will be improved by aligning cognitive incongruence or misfit between leaders and followers, or among different disciplinary members. Established discipline guides to manage technological procedures but does not appropriately lead to align value consciousness or appreciation among heterogeneous members. Methodology should be developed in enhancing collaborative work and in improving operational flexibility and organizational resilience.

BACKGROUND

Organizational effectiveness is assessed by the degree of appreciation on organizational context as well as by the degree of job skill and disciplinary knowledge related. The relationship of job skill with task performance has been studied, but there are few studies on contextual performance and very few on human performance, which is influenced by the practical process in the workplace or by the recursive learning process in organizational environment. Human performance emerges in the recursive information processing of organizational learning in which cognition and action are inseparably coupled with each other. Hence, whole system approaches or holistic approaches to identify what cognition-action coupling process of organizational learning necessitates for the evaluation of a human performance by treating a human as a social being. The theories on the continuous or recursive process of organizational learning in the field we have referred to in our surveys are appreciation system theory (Vickers, 1965, 1983), soft systems methodology (Checkland, 1993; Checkland & Holwell, 1998; Checkland & Scholes, 1999), knowledge creation process (Nonaka & Konno, 1998; Nonaka & Nishigihi, 2001)The author found later on that the essence of Vickers' theory was quite similar to the "pure experience" that emerged in practical field and that was the essence of Nishida's (1911) theory. Appreciation processes in the field, in addition to the already mentioned theories such as seven stages of internal and external thinking process in soft system methodology (Checkland, 1993), four stages of socialization, externalization, combination, and internalization in knowledge creation process (Nonaka, & Kanno, 1998), and many other processes, were reported by researchers in the past (Deming, 1982; Gomez, 1999; Kaplan & Norton, 2006; Maruyama, 1992, 1994; Mingers & Gill, 1997; Ulrich & Probst, 1991; Probst & Gomez, 1992). In this chapter, the degree of appreciation rather than the recursive process of appreciation on working conditions and organizational context is focused to assess human performance in the level of individuals and collectives.

MAIN FOCUS OF THE CHAPTER

The main focus of this chapter is on the relation of the degree of appreciation with organizational performance of the participants in a health care organization. Cognitive or value congruence among individuals and collectives leads to increased organizational effectiveness, while value incongruence leads to increased withdrawal behaviors, such as absenteeism and turnover. In this chapter, some empirical evidences provided are to positively identify significant effects of appreciation degree on organizational performance and to compare multiple causal relationships between the higher and the lower groups of appreciation.

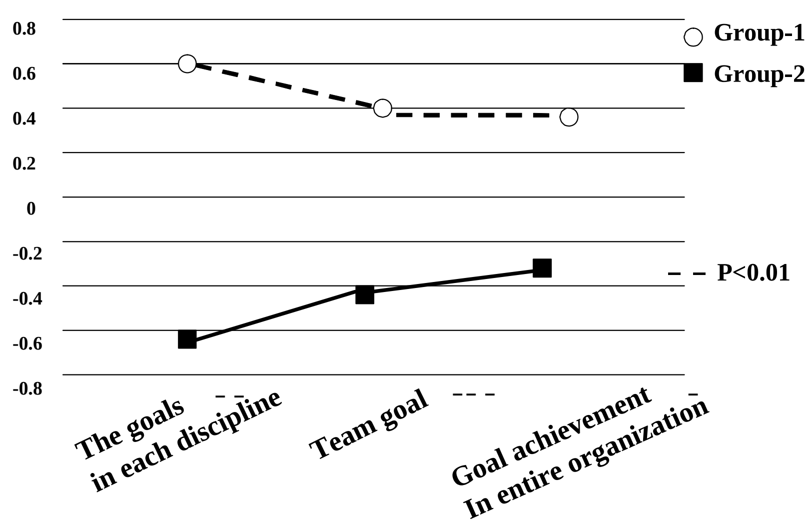
A total of 175 nurses engaging in a hospital, their average age and SD 32.1 ± 9.1, were investigated for the purpose of identifying the relation between organizational performance and individual appreciation of organizational environment. All the subjects were asked to respond to the self-administered questionnaires on work environment, learning type, fairness, and job satisfaction prepared for the survey using a four-point Likert-type scale. The most important indicators in evaluating performance of hospital nurses are performance reliability and team coherence or reciprocity. Performance reliability was measured by the common performance conditions (CPCs) developed by Hollnagel (1993, 1998). Human error is predicted by using performance reliability measured by nine items of work environmental conditions, which are called the CPCs, asking how do you cognize each working condition in operational, strategic, and normative levels; namely, asking how do you cognize the (1) adequacy of an organization, (2) working conditions, (3) adequacy of Man-Machine Interface and operational support, (4) availability of procedures/plans, (5) number of simultaneous goals, (6) available time, (7) time of day/circadian rhythm, (8) adequacy of training and preparation, and (9) crew collaboration quality, which are developed by Hollnagel (1998). Subjects are requested to select one of four answers: very efficient, efficient, inefficient, deficient.

Team coherence was measured by using the team-member exchange construct (TMX) developed by Seers (1989) and Seers, Petty, and Cashman (1995), which was designed to address an employee's exchange relationship to the peer group as a team. The TMX assesses the reciprocity among team members with respect to the members' contribution of ideas, feedback, and assistance to other members, made up of 34 questionnaire items with a five-point scale of -2, -1, 0, +1, and +2.

CLASSIFICATION OF APPRECIATION LEVEL

Subjects consisting of 175 hospital nurses were classified into two groups: higher and lower groups in the appreciation level of disciplinary goals in individual goals, team goals, and goal achievements of the organization. As shown in Figure 1, significant differences between the two groups was observed in terms of the goals in the discipline, team goal, and goal achievement in the organization (Saito, Murakami, Nishiguchi & Seki, 2006). In the higher group, Group 1, the

Figure 1. Comparison of appreciation on disciplinary goal, team goals and goal achievement of organization



degree of appreciation in achieving disciplinary goals appeared slightly higher than the appreciative degrees in team goals and organizational goals, while in the lower group, the degree of disciplinary goals appeared slightly lower than the degree of appreciation in team goals and organizational goals. These results suggest that development of individual disciplinary knowledge and skills seems important for professional workers like nurses in the higher level of appreciation, while nurses in the lower level of appreciation seem uninterested and nonchalant in individual development as well as achievement of team and organizational goals.

COMPARISON OF TEAM COHERENCE, AND PERFORMANCE RELIABILITY BETWEEN TWO GROUPS

Organizational performances such as team coherence or team reciprocity and performance reliability as two important indicators assessing organizational performance were compared between two groups clustered with appreciation on goals.

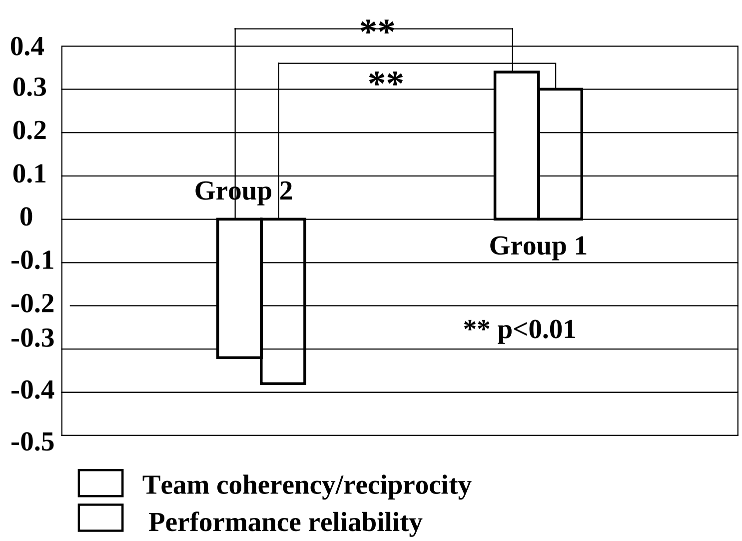
As shown in Figure 2, organizational performances represented by team coherence and performance reliability significantly ($P<0.01$) differed between two groups of the higher and the lower groups in appreciation (i.e., Group 1 and Group 2). Nurses in Group 1 are expected to provide care in security and safety more than the ones in Group 2. Incidence experience and accidental occurrence caused by human error is predicted by performance reliability according to the CREAM (Cognitive Reliability and Error Analysis Method) (Hollnagel, 1998). Average weighing factors of error occurrence for four control modes, such as scrambled, opportunistic, tactical, and strategic, are 2.3E+01, 7.5E+00, 1.9E+00, and 9.4E-01.

Nurses in Group 1 who marked higher factor loadings in team reciprocity are able to maintain better interpersonal relations among team members than the ones in Group 2. Group 1 might be engaging in a recursive process of organizational learning for keeping team coherence and performance reliability higher. Individual and organizational performance in Group 1 might be leveraged by communication through organizational learning and might be more effective and cooperative than in Group 2.

COMPARISON OF ORGANIZATIONAL PERFORMANCES, FAIRNESS AND JOB SATISFACTION

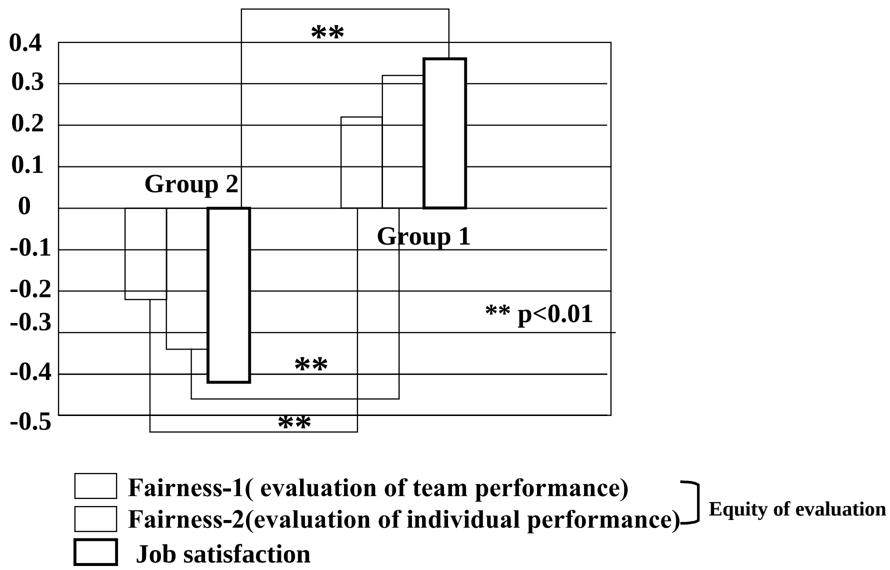
Organizational performances represented by fairness-1 and -2, and job satisfaction significantly ($P<0.01$) differed between two groups of the higher and the lower

Figure 2. Comparison of organizational performance of team coherency and performance reliability between higher and the lower in appreciation of organizational goal



A

Figure 3. Comparison of organizational performance of fairness -1 and -2 and job satisfaction between the higher and the lower in appreciation of organizational goal



degrees in appreciation of team care goals and organizational goals, as shown in Figure 3. Fairness-1 signifies how subjects feel about team performance evaluation, and fairness-2 means how subjects feel about individual effort-demand fairness (Janssen, 2001; Siegrist, 1996). Job satisfaction represented by satisfaction about the work allotted, meaning of work, and work pride significantly differed between the two groups, the higher and the lower degree of appreciation. The result suggests that the degree of appreciation on organizational context plays a significant role in improving the feeling of fairness and job satisfaction.

COMPARISON OF MULTIPLE CAUSAL RELATIONSHIP AMONG WORK DEMAND, ORGANIZATIONAL ENVIRONMENT, AND ORGANIZATIONAL PERFORMANCE

Causal relationship in the lower group was constructed by the factors representing physical and extrinsic factors, such as work time and frequency of patient visit for caring, while causal relationship in the higher group was constructed by the factors representing intrinsic factors, such as work discretion, mutual understanding, and work cooperation. As shown in Figure 4, the meanings of structural variables on work demands, organizational environment, and fairness evaluation were different between the two groups. As to fairness evaluation, work cooperation representing work demand and team work atmosphere representing organizational environment in Group 1 of higher appreciation linked to criterion variable of fairness in the evaluation of individual effort, while in Group 2 of lower appreciation, working time, and time for patient care representing work demand and organizational environment linked to criterion variable of overall evaluation by the hospital. These linkages were testified by good fit index (GFI=0.968, AGFI=0.903, RMR=0.024 in Group 1; GFI=0.931, AGFI=0.825, RMR=0.018 in Group 2). Direction and strength in causation among three variables such as work demand, organizational environment, and fairness as a criterion variable and the meaning of the structural variables, were different between the two groups.

As shown in Figure 5, in the causal linkages among work demand, organizational environment, and job satisfaction as a criterion variable, the meanings of each structural variable and criterion variable were different between the two groups, work cooperation representing work demand, understanding among staff representing organizational environment and pride for job representing job satisfaction in Group 1, while in Group 2, the meanings of all the structural variables

Figure 4. Casual relationship among environment, work demand, and fairness of evaluation in Group 1 and Group 2, the higher and the lower groups in work appreciation

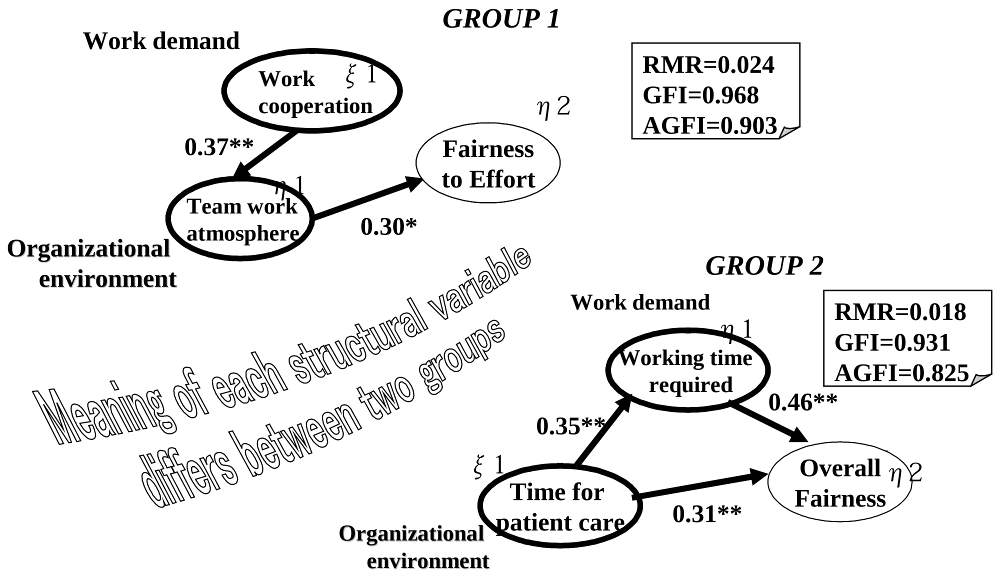
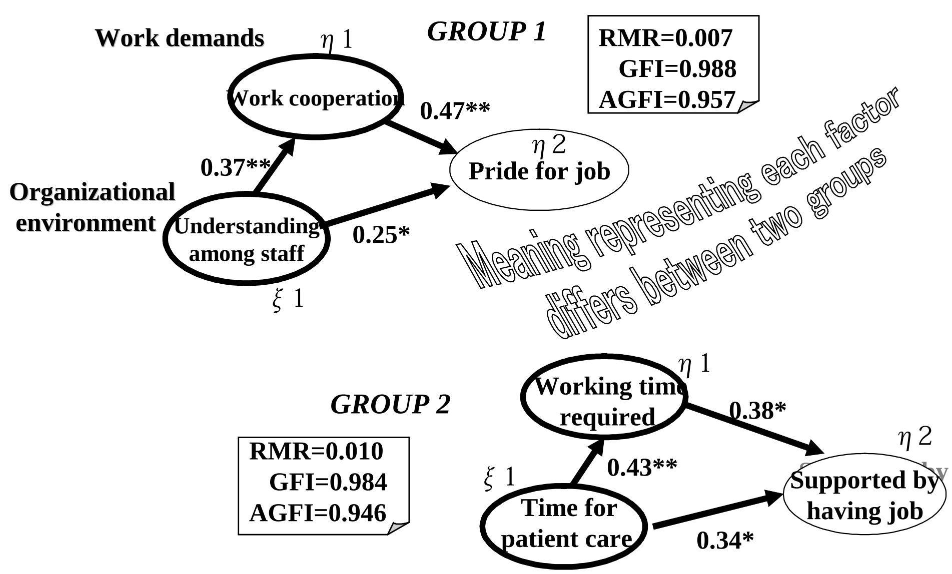


Figure 5. Casual relationship among environment, work demand, and job satisfaction in Group 1 and Group 2, the higher and the lower groups in work appreciation



were physical, such as working time representing work demand, time required for patient care representing organizational environment, and satisfaction by having job. Both models of causation were testified by good ft index (GFI=0.988, AGFI=0.957, RMR=0.007 in Group 1; GFI=0.984, AGFI=0.946, RMR=0.010 in Group 2).

FUTURE TRENDS

Limitation of this chapter is classification of appreciation on organizational context in terms of goals in each discipline, team goals, and organizational goals. Organizational contextual factors other than the factors adopted in this chapter exist in the workplace. Typology and classification should be tried by including management strategies and their implementation processes, and also members' characteristics (i.e., knowledge, skills, attitudes, and others (the KSAOs). Appreciation in the workplace may be different by the degree of cognitive alignment between individuals and collectives, and also alignment of cognition and action on the individual level. Individual cognition on job responsibility and reciprocity among superiors, peers, and other stakeholders give a critical effect on job performance. Individual job competences might play crucial roles as moderators or mediators between organizational constraints as antecedents and consequences of organizational performances. Job competence of emotional regulation in the workplace is one of the influential factors as far as we have surveyed. Future studies are expected to identify interactive relations between individual and collective competences. Integrated value management approaches, whole system approaches, or holistic approaches are left for future study.

CONCLUSION

1. The degree of appreciation in an organizational context gave significant effects on team coherence, performance reliability, fairness perception, and job satisfaction. The higher degree of appreciation on organizational environment played a significant role in keeping a higher perception of team coherence/team reciprocity, performance reliability, fairness both in individual and team performance, and job satisfaction.
2. Multiple causation between the higher and lower groups of appreciation differed in the intensity and direction of causation among three variables of work demand, organizational environment, and fairness perception or job satisfaction as shown in Figure 4 and Figure 5. In the higher group of appreciation, work cooperation and understanding among staff members led to admitting equity of evaluation and keeping pride for job, while in the lower group of appreciation, working time and time required for patient care are strongly related to overall or vague feeling of fairness and acceptance feeling as a job. The result suggests that having higher appreciation signifies substantiality a cognitive world, while keeping lower appreciation signifies just following daily work without any development.

REFERENCES

Checkland, P. (1993). *Systems thinking, systems practices*. John Wiley and Sons, Inc.

Checkland, P., & Holwell, S. (1998). *Information, systems, and information systems: Making sense of the field*. John Wiley and Sons, Inc.

Checkland, P., & Scholes, J. (1999). *Soft systems methodology in action*. John Wiley and Sons, Inc.

Deming, E.W. (1982). *Quality, productivity and competitive positions*. Cambridge, MA: MIT Press.

Flood, R. (1993). *Beyond TQM*. John Wiley and Sons Inc.

Gomez, P. (1999). Integrated value management.

Hollnagel, E. (1993). *Human reliability analysis: Context and control*. Academic Press.

Hollnagel, E. (1998). *Cognitive reliability and error analysis method*. Elsevier.

Janssen, O. (2001). Fairness perception as a moderator in the curvilinear relationships between job demands, job performance and job satisfaction. *Academy of Management Journal, 44*(5), 1039–1050.

Kaplan, R.S., & Norton, D.P. (2006). *Alignment: Using the balance scorecard to create corporate synergies* Harvard Business School Press.

Maruyama, M. (1992). *Context and Complexity: Cultivating contextual understanding.* Springer-Verlag.

Maruyama, M. (1994). *Mindscapes in management: Use of individual differences in multicultural management*. Dartmouth Publishing Co.

Mingers, J., & Gill, A. (1997). *Multimethodology: The theory and practice of combining management science methodologies*. John Wiley and Sons, Inc.

Nishida, K. (1911). *An inquiry into the good*. Tokyo: Iwanami Book Publishing Co.

Nonaka, I., & Konno, N. (1998). The concept of "ca": Building a foundation for knowledge creation. *California Management Review, 40*(3), 1–15.

Nonaka, I., & Nishiguchi, T. (2001). *Knowledge emergence: Social, technical and evolutionary dimensions of knowledge creation*. Oxford University Press.

Probst & Gomez, P. (1992). Thinking in networks to avoid pitfalls of managerial thinking. In M. Maruyama (Ed.), *Context and complexity: Cultivating contextual understanding*. Springer-Verlag.

Saito, M., Murakami, G., Nishiguchi, H., & Seki, H. (2006). *Job cognition-action causation in hospital nurses for measurement of team care performance*. Proceedings of the 9th International Congress of Behavioral Medicine (ICBM), Bangkok, Thailand.

Seers, A. (1989). Team member exchange quality: A new construct for role-making research. *Organizational Behavior and Human Decision Processes, 43*, 118–135.

Seers, A., Petty, M.M., & Cashman, J.F. (1995). Team member exchange under team and traditional management: A naturally occurring quasi-experience. *Group & Organization Management, 20*, 18–38.

Siegrist, J. (1996). Adverse health effects of high-effort/low-reward conditions. *Journal of Occupational Health Psychology, 1*(1), 27–41.

Ulrich, H., & Probst, G.J.H. (1991). *Anleitung zun ganzheitlichen denken und handeln verlag Paul Haupt* (in German).

Vickers, G. (1965). *The art of judgment*. Harper and Row.

Vickers, G. (1983). *Human systems are different*. Harper and Row.

KEY TERMS

Appreciation: Value consciousness or understanding in organizational environments in which subjects engage. The degree of appreciation in organizational contexts has an effect on individual and organizational performances. In this chapter, appreciation was measured by clustering subjects into two groups, the higher and lower groups, in terms of the degree of appreciation on individual disciplinary goals, team goals, and organizational goals.

Fairness: Fairness in this chapter is defined in two aspects: fairness-1 signifies how subjects feel about team performance evaluation; and fairness-2 means how subjects feel about individual effort-demand fairness (Janssen, 2001; Siegrist, 1996) by preparing 14 questionnaire items with a four-point scale: totally disagree to totally agree.

Job Satisfaction: Measured by preparing six questionnaire items representing satisfactions about the work allotted, meaning of work, and work pride with a four-point scale: strongly disagree to strongly agree.

Performance Reliability: In this chapter, concept and methodology of performance reliability developed by Hollnagel (1998) were applied in predicting human error during work, measured by using nine common performance conditions (the CPCs) with a four-point scale: very efficient, efficient, inefficient, and deficient. Improved reliability and reduced reliability are also measurable by using the CPCs.

Team Coherence: Measured by applying the TMX (Team Member Exchange Quality) developed by Seers (1989). Team coherence was evaluated in this chapter to define mutual support, group cohesiveness, trust, responsibility for team goals, and team reciprocity as defined by the developer.

Arrhythmia Detection and Classification Using Wavelet and ICA

Vahid R. Sabzevari
Azad University of Mashhad, Iran

Asad Azemi
The Pennsylvania State University, USA

Morteza Khademi
Ferdowsi University of Mashhad, Iran

Hossein Gholizade
Ferdowsi University of Mashhad, Iran

Armin Kiani
Ferdowsi University of Mashhad, Iran

Zeinab S. Dastgheib
Ferdowsi University of Mashhad, Iran

INTRODUCTION

The classification of an electrocardiogram (ECG) into different pathophysiological disease categories is a complex pattern recognition task. Computer-based classifications of the ECG can achieve high accuracy and offer the potential of an affordable mass-screening for cardiac abnormalities. Successful classification is achieved by finding the characteristic shapes of the ECG that discriminate effectively between the required diagnostic categories. Conventionally, a typical heart beat is identified from the ECG, and the component waves of the QRS, T, and possibly P waves are characterized using measurements such as magnitude, duration, and area. Datasets that are used for heart diseases involve different features. Some of them are based on laboratory experiments, while others include clinical symptoms. However, one of the most popular and useful databases is the MIT-BIH (http://physionet.fri.uni-lj.si/physiobank/database/mitdb/). Researchers have used this database to test their various algorithms for arrhythmia detection and classification. Two of the most popular methods are artificial neural networks (ANNs) and wavelet transforms, and their variations. For example, Lee (1990) classified three types of cardiac arrhythmias with accuracies of 99.55%, 97.75%, 57.1%, respectively using ANNs. Chi and Jaberi (1992), using triple neural networks, classified ventricular arrhythmia with the average accuracy of 95.1%. Karlik and Ozbey (1996) were able to classify 10 types of arrhythmia with average accuracy of 95%. Hu (1997), using self-organizer neural networks and the Linear Vector Quantization, was able to classify heart beats with average accuracy of 91.3% and 90.3%, respectively. Gholam Hosseini, Rainolds, and Powers (2001) implemented a multistage neural network with two MLPs, and were able to classify five kinds of arrhythmia. At the first stage, they achieved an average accuracy of 81.8%, and at the second stage, 88.3%. The algorithms developed in these works are based on analyzing the ECG signals, which requires a large simulation time.

Selected examples of using wavelet transforms and their variations for arrhythmia classification include the following. Yang, Hu, and Shyu (1997), using dyadic wavelets to extract features, and Kohonen self-organizing neural networks for classification, were able to obtain an average precision of 97.77% for heart disease diagnosis. Dokur, Olmez, and Yazgan (1999) used discrete wavelet transforms to classify ten types of arrhythmias with a precision of 97%. De Chazal, Celler, and Reilly (2000), using a set of 500 records with 345 abnormal cases, classified arrhythmias by using 15 feature sets of three Daubechies wavelets decomposition level and reached the maximum precision of 74.2%. Finally, Dokur and Olmez (2001), using wavelet transforms and ANNs trained by genetic algorithms back propagation, were able to classify 10 types of arrhythmia with a precision of 96%.

The goal of this article is to optimize the feature extraction process by using ICA and wavelet transform, apply the obtained set to several different machine learning schemes, and compare their performances. The article is structured as follows. Section 2.0 describes our proposed method for cardiac arrhythmias detection. Section 3.0 covers an overview of different classifier types that were used in this work. Sections 4.0 and 5.0 summarize our simulation scheme and results. Finally, section 6.0 presents the concluding remarks.

PROPOSED METHOD

Figure 1 presents the block diagram of the proposed detection and classification process. First, the appropriate components of the ECG signal are obtained by using the ICA algorithm. Next, these components are used to calculate the coefficients of the Daubechies wavelets. Based on this step, proper features are selected and fed into the classifier. For comparison purposes, three different machine learning methods have been implemented—namely, radial basic function (RBF), multilayer perceptrons (MLPs), and K-nearest neighbor (KNN).

Next, we will present an overview of the ICA and wavelet transform.

Independent Component Analysis with a Time Structure Method

To rigorously define ICA, we can use a statistical "latent variables" model. We observe n random variables $x_1,\ldots,x_n$, which are modeled as linear combination of n random variables $s_1,\ldots,s_n$:

$$x_i = a_{i1}S_1 + a_{i2}S_2 + \ldots + a_{in}S_n \quad i = 1,\ldots,n \qquad (1)$$

Where a_{ij}, $i,j=1, \ldots, n$ are some real coefficients. By definition, the s_i are statistically and mutually independent.

This is the basic ICA model. The ICA model is a generative model, which means that it describes how the observed data are generated by a process of mixing the components s_i. The independent components s_i (often abbreviated as ICs) are latent variables, meaning that they cannot be directly observed. Also, the mixing coefficients a_{ij} are assumed to be unknown. All we observe are the random variables x_i, and we must estimate both the mixing coefficients a_{ij} and the ICs s_i using the x_i.

Here, we have dropped the time index t, because in this basic ICA model, we assume that each mixture x_i, as well as each independent component s_i, is a random variable, instead of a proper time signal or time series.

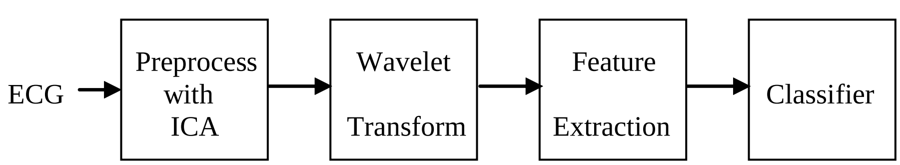
If the independent components (ICs) are time signals, the situation is quite different. In many applications, however, what is mixed is not random variables, but time signals, or time series. This is in contrast to the basic ICA model in which the samples of x have no particular order; we could shuffle them in any way we like, and this would have no effect on the validity of the model. In fact, if the ICs are time signals, they may contain much more structure than simple random variables.

In this research, we consider the estimation of the ICA model when the ICs are time signals, $s_i(t), t=1,\ldots,T$, where t is the time index. Here, t has a more precise meaning, since it defines an order between the ICs. The model is then expressed by:

$$x(t) = As(t) \qquad (2)$$

where A is assumed to be square as usual and the ICs are of course independent. We shall make some as-

Figure1. Block diagram of the proposed detection-classification system



sumptions on the time structure of the ICs that allow for the estimation of the model. First, we shall assume that the ICs have different autocovariances (in particular, they are all different from zero). Second, we shall consider the case where the variances of the ICs are nonstationary, as described by Hyvarinen, Karhunen, and Oja (2001).

In this article, we have implemented the "Fast ICA" algorithm to calculate ICs, similar to the work done by Hurri, Gavert, Sarela, and Hyvarinen (2005). According to the previous research done in this area, Rieta, Castells, Sánchez, Zarzoso, and Millet (2004), and Castells, Rieta, Millet, and Zarzoso (2005), at least three factors make ICA suitable for ECG signal analysis. These factors are as follow:

1. A trial and ventricular activity (AA, VA) are generated by sources of independent bioelectric activity;
2. AA and VA present non-Gaussian distributions; and
3. The generation of the surface ECG potentials from the cardioelectric sources can be regarded as a narrow-band linear propagation process.

Figure 2 shows results of the ICA processing and the original signal, which was generated by two independent sources of atrial and ventricular electrical activity (in such a ventricular arrhythmia). As shown in this figure, ICA with time structure successfully described these two sources that are characterized by "o" and "+" symbols in this figure.

Wavelet Analysis

Wavelet Analysis of a signal consists of breaking up a signal into a translated (shifted) and dilated (scaled) version of a reference (mother) wavelet. A wavelet is a signal of effectively limited duration that has an average value of zero. In determining the wavelet (decomposition) coefficients of a signal, the correlation of the mother wavelet at different translations and dilations with the signal is computed. Hence, the wavelet coefficients represent measures of similarity of the local shape of the signal to the mother wavelet under different translations and dilations. We utilize the feature extraction ability of this analysis for ECG. In this work, we have used Daubechies wavelets; this choice is explained in the Simulation section.

CLASSIFIERS

The results obtained from the previous section were fed into three different machine-learning algorithms—namely, (Multilayer Perceptron) MLP, (Radial Basis Function) RBF, and K-nearest neighbor. An overview of each of these methods follows.

Multilayer Perceptron

The most widely used neural classifier today is a Multilayer Perceptron network, which has also been extensively analyzed, and for which many learning algorithms have been developed. The MLP belongs to the class of supervised neural networks. The network has a simple interpretation as a form of an input-output model, with the weights and thresholds (biases), the free parameters of the model. Such networks can model functions of almost arbitrary complexity, with the number of layers, and the number of units in each layer, determining the function complexity. Multilayer networks use a variety of learning techniques, the most popular being back-propagation. Here, the output values are compared with the correct answer to compute the value of some predefined error-function.

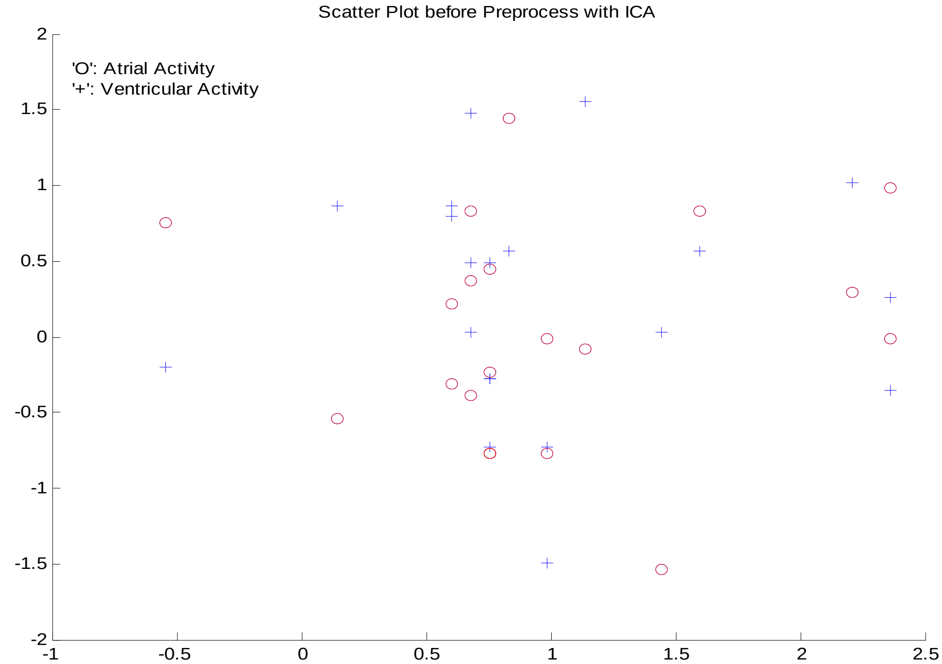
Radial Basis Function

Radial Basis Function Networks used for pattern classification are based on Cover's theorem on the separability of patterns. It contains an input layer, a hidden layer with nonlinear activation functions and an output layer with linear activation functions. The RBF network is a popular alternative to the MLP, which, although it is not as well-suited to larger applications, can offer advantages over the MLP in some applications. The RBF network has a similar form to the MLP, in that it is a multilayer, feed-forward network. However, unlike the MLP, the hidden units in the RBF are different from the units in the input and output layers: they contain the "Radial Basis Function," a statistical transformation based on a Gaussian distribution from which the neural network's name is derived.

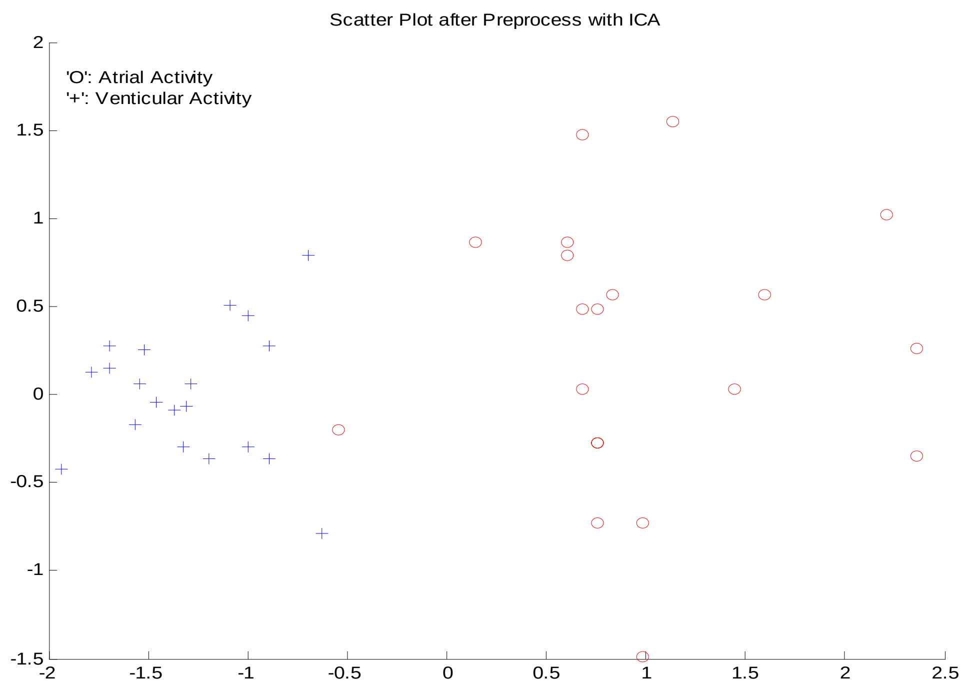
K-Nearest Neighbor Classifier

The nearest neighbor classifier algorithm is a method for classifying phenomena based upon observable features. In this algorithm, each feature is assigned a

Figure 2. Scatter plots (a) before denotation of independent sources by ICA; and (b) after denotation by Matlab ICA pack



(a)



(b)

dimension to form a multidimensional feature space. A training set of objects, with a priori-known class are processed by feature extraction and plotted within the multidimensional feature space. The offsets in each dimension are referred to as the feature vector. This is the training or learning stage. The geometric distance is computed between the new feature vector, and each a priori-feature vector from the training set. The shortest distance thus computed is to the nearest neighbor. The a priori class of the nearest neighbor is now assigned to the phenomena to be classified.

The simulation results of using the aforementioned profilers showed that the MLP provides the best classification. Details of the MLP classifier are presented in the next section. Table 1 summarizes the comparative simulation results:

Table 1. A comparison of the different profiler performance

Method	Overall Accuracy
RBF $(s = 0.3)$	70.80%
KNN (k=5)	84.80%
MLP (3 layers)	98.40%

SIMULATION

For the simulation part, we have used MATLAB's Neural Networks and Wavelet toolboxes, along with its Fast ICA package for pattern recognition and classification.

Five data sets, one normal and four abnormal, with modified lead II (MLII) and lead V1 from MIT-BIH were used in training and testing stages. The first set was from the NSR (Normal Sinus Rhythm), the second set from PVC (Premature Ventricular Contraction), the third set from LBBB (Left Bundle Branch Block), the fourth set from RBBB (Right Bundle Branch Block) and finally, the fifth set from P (Paced Beat) data bases. Dilation coefficients of Daubechies wavelets levels 6, 7, 8 of MLII and dilation coefficient levels 6, 7 of V1 samples after denotation by ICA were implemented as features. Next, the wavelet coefficients were normalized, and then fed into MLP, and the classification results were obtained. Table 2 illustrates the data distribution, and the number of records used for this part of the simulation.

The MLP, which is implemented in this article, is a three-layered neural network. The optimum topology was obtained with the following structure: five input neurons, eight hidden neurons, five output neurons, and the Levenberg-Marquardt as the training algorithm. The desired error was set at 0.01. The system was tested for 1–10 hidden neurons, and four different discriminator functions (activation functions).

Table 2. Data set descriptions and numbers used in the simulation

Description	No. of beats used in testing	No. of beats used in training	No. of records in MIT-BIH	Class No.
Normal Beat	50	100	101,105	1
PVC	50	100	200,213	2
LBBB	50	100	207,214	3
RBBB	50	100	118,212	4
P	50	100	107,217	5
	250	500	Total =750	

SIMULATION RESULTS

Table 3 illustrates the comparison of different activation functions that have been used in the MLP network that was described in the previous section. An entry such as "tansig-logsig" means that "tansig" is the activation function of the hidden neuron, and "logsig" is the activation function of the output layer. As the table indicates, optimum results can be obtained with the "tansign-tansig" activation functions with an overall accuracy of 98.4%, during testing. This is an improvement in comparison to the ones obtained in Shyu, Wu, and Hu (2004).

The performances of the MLP classifier are also expressed in terms of specificity, sensitivity, and overall accuracy. The specificity is defined as the fraction of correctly classified normal rhythms. The sensitivity of an arrhythmia is defined as the fraction of that specific arrhythmia correctly classified. The overall accuracy is the fraction of the total ECG beats correctly classified. The specificity and sensitivities of different ECG beat types are summarized in Table 4 for comparison.

Table 3. The overall accuracy in testing of MLP for different activation function (100% accuracy in training)

MLP Accuracy (%)	Activation functions
98.40	tansig-tansig
91.20	logsig-tansig
90.00	tansig-logsig
92.80	logsig-logsig

Table 4. Classification results of the proposed method

100.00	Specificity (%)	
98.00	PVC	Sensitivity (%)
96.00	LBBB	
98.00	RBBB	
100.00	P	
98.40	Accuracy (%)	

CONCLUSION

In this article, we have shown a new method for cardiac arrhythmias detection and classification, based on using Independent Component Analysis and wavelet transforms to extract important features. The extracted features were then fed to an MLP with the Levenberg-Marquardt back propagation algorithm. For the simulation part, 750 instances of MIT-BIH database from five disease categories were used and a 98.4% accuracy was obtained, which is an improvement with respect to similar works

According to the simulations, the type of activation function and the number of hidden nodes are the important factors in the topology of neural networks. We note that for extending these results, we need to use a larger number of samples from the database.

REFERENCES

Castells, F., Rieta, J. J., Millet, J., & Zarzoso, V. (2005). Spatiotemporal blind source separation approach to atrial activity estimation in atrial tachyarrhythmias. *IEEE Trans. on Biomedical Engineering, 52*(2).

Chi, Z., & Jaberi, M. A. (1992). *Identification of supraventricular and ventricular arrhythmias using a combination of 3NNs*. Department of Electrical Engineering, University of Sydney.

De Chazal, P., Celler, B. G., & Reilly, R. B. (2000). Using wavelet coefficients for the classification of the electrocardiogram. In *Proceedings of the 22nd Annual EMBS International Conference* (pp. 64-67). Chicago.

Dokur, Z., & Olmez, T. (2001). ECG beat classification by a novel hybrid neural network. *Computer Methods and Programs in Biomedicine, 66,* 167–181.

Dokur, Z., Olmez, T., & Yazgan, E. (1999). Comparison of discrete wavelet and fourier transforms for ECG beat classification. *IEEE Electronic Letters, 35*(18), 1502-1504.

Gholam Hosseini, H., Rainolds, K. J., & Powers, D. (2001). A multi-Stage NN. Classifiers for ECG events. In *Proceedings of Annual International Conference of the IEEE Eng. of Med. and Biology Society*.

Hurri, J., Gavert, J., Sarela, J., & Hyvarinen, A. (2005). *The Fastica package for Matlab*. Retrieved February 4, 2008, from http://www.cis.hut.fi/projects/ica/fastica

Hu, Y. H. (1997). Customization of ECG beat classifiers developed using SOM&LVQ. In *Proceedings of IEEE* (pp. 813-814).

Hyvarinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis*. Wiley-Interscience.

Karlik, B., & Ozbey, Yu. (1996). A new approach for arrhythmia classification. In *Proceedings of Annual International Conference of the IEEE Eng. of Med. & Biology Society*.

Lee, S. C. (1990). Using a translation-invariant neural network to diagnose heart arrhythmia. In *Proceedings of Annual International Conference of the IEEE Eng. of Med. and Biology Society* (Vol. 12, pp. 2025-2029).

Rieta, J. J., Castells, F., Sánchez, C., Zarzoso, V. & Millet, J. (2004). Atrial activity extraction for atrial fibrillation analysis using blind source separation. *IEEE Transactions on Biomedical Engineering, 51*(7), 1176-1186.

Shyu, L. -Y., Wu, Y. -H., & Hu, W. (2004). Using wavelet transform and fuzzy neural network for VPC detection from the holter ECG. *IEEE Transactions on Biomedical Engineering, 51*(7), 1269-1273.

Yang, M. -Y., Hu, W. -C., & Shyu, L. -Y. (1997). ECG events detection and classification using wavelet and neural networks. In *Proceedings of 19th International Conference IEEE/EMBS* (pp. 289-292). Chicago.

KEY TERMS

Artificial Neural Network (ANN): An interconnected group of artificial neurons that uses a mathematical model, or computational model, for information processing, based on a connectionist approach to computation.

Cardiac Arrhythmia: A group of conditions in which the muscle contraction of the heart is irregular, or is faster or slower than normal.

Independent Component Analysis: A computational method for separating a multivariate signal into additive subcomponents supposing the mutual statistical independence of the non-Gaussian source signals.

K-Nearest Neighbor Algorithm (K-NN): A method for classifying objects based on closest training examples in the feature space.

Multilayer Perceptron: A class of networks consists of multiple layers of computational units, usually interconnected in a feed-forward way.

Radial Basis Function (RBF): A real-valued function whose value depends only on the distance from the origin. In artificial neural networks, radial basis functions are utilized as activation functions.

Wavelet Transform: Representation of a signal in terms of a finite length or fast decaying oscillating waveform (known as the mother wavelet). This waveform is scaled and translated to match the input signal.

Assessing the Information Content of Microarray Time Series

E. Yang
Rutgers University, USA

I.P. Androulakis
Rutgers University, USA

INTRODUCTION

While the rise of microarray**s** has heralded a new era in molecular biology with its ability to measure the expression level of thousands of genes at once, the usefulness of microarray**s** is exigent upon the ability to obtain accurate gene expression data for the individual genes (Bowtell, 1999; Brown & Botstein, 1999; Cheung, Morley, Aguilar, Massimi, Kucherlapati, & Childs, 1999). However, there has been significant criticism as to how meaningful the information derived via microarrays is. In cases where one has attempted to find genes that correlated to types of cancer or survival rate, it was found that different analysis techniques would often times yield radically different set of genes, calling into question the validity of the overall experiment itself (Dupuy & Simon, 2007). It is our contention that part of the problem associated with microarrays is that there does not exist a coherent method for dealing with data quality, and if a coherent method for dealing with data quality existed, many of the criticisms of microarrays could be addressed.

BACKGROUND

"Fishing Expedition" is normally used in a negative connotation in the legal field. The negativity has carried over to the scientific field and describes an experiment in which the researcher does not know precisely what one is looking for. However, the promise of microarrays is the fact that they allow for just such experiments, and coupled with different algorithms allow for a data driven approach to science (Nakai & Vert, 2002). By identifying the possible targets of gene regulation, researchers can then formulate more specific experiments to validate such a hypothesis. While the technology behind microarrays consistently advances in terms of the density of microarrays as well as the repeatability between each individual microarray, there still exists the major issue of noise. Whilst the technological improvements themselves are able to minimize the technical noise, there still exists significant *biological noise* which for complex multitissue organisms cannot be easily overcome with technology. For instance, despite the standardization of rat/mice lines, there still exists significant variation in any reading taken from a population of animals. Due to this noise, it is difficult to identify the genes which actually respond to a given treatment, and those genes whose fluctuations are due to random noise.

Therefore the most important algorithms for the processing of microarray data are those that select meaningful genes from the thousands that are measured via the microarray. The most common metric used by these algorithms is that of *statistical significance* (Smyth, Yang, & Speed, 2003). There is one caveat with the use of statistical significance primarily in the fact that not all genes that are statistically significance are biologically relevant. The set of biologically relevant genes is dependent wholly upon the biological ground truth, whilst the set of statistically significant genes are dependent upon the number of replicates, the quality of the microarray platform, inherent SNR, as well as biological significance. This is a recognized problem which has been addressed by many researchers as a *feature selection* problem under the assumption that the set of biologically relevant genes ought to be able to work as classifiers between the different states being tested (Wu, 2005).

While not all genes that are statistically significant are biologically relevant and vice versa, there does exist a tendency for biologically relevant genes to be statistically significant as well. Therefore, by selecting statistically significant genes, one increases the likelihood of identifying biologically relevant genes as well. However for one to have confidence in these initial results, care must be taken in the selection of statistically significant genes, paying special attention to normalization and the setting of statistically significant cutoffs.

STATISTICAL SELECTION OF GENES FROM MICROARRAYS

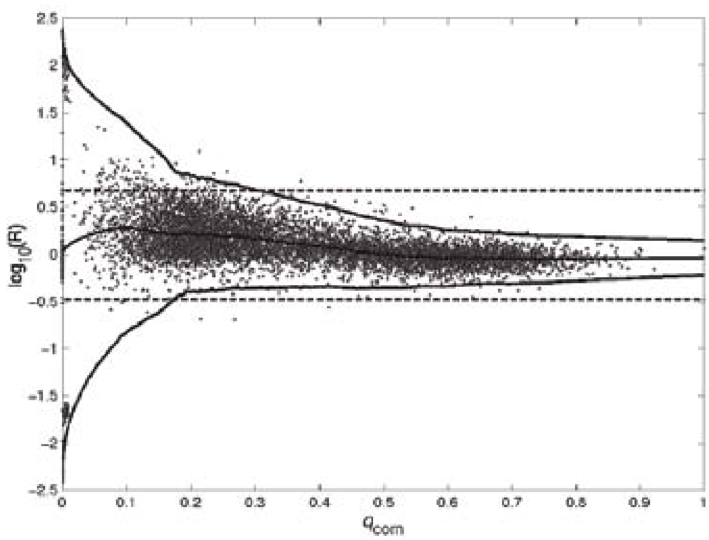
Normalization

The use of normalization is important because many of the selection algorithms that look for over/under expressed genes in a two state experiment base their assumption off of the fact that the initial distribution log is normalized and compute their confidence intervals according to that distribution. Therefore by transforming the data so it does conform to the log-normal distribution, one is able to use standard statistical tests such as the t-test to ascertain whether or not the variations are due to noise or due to some intrinsic change in the expression level of the gene.

One of the challenges with analyzing microarray data is the problem of translating the recorded intensity level obtained by the detection equipment and determining the true expression value of the given probe. Generally, for genes which show a high intensity value, there exists a good correlation between the intensity between the two samples follows a roughly linear trend. However, at low intensity levels, the correlation between the two samples deviates from this (see Figure 1). The justification for the majority of genes being linearly correlated is that under most situations, only a small fraction of genes are responding to the overall treatment and even with the addition of noise, they should be consistent over multiple chips in *temporal gene expression* experiments. The small fraction of genes that do deviate from this linear relationship are then the ones that deviate by a given statistically significant level. Techniques such as the LOESS, dCHIP, and PDNN (Cleveland, 1979; Li & Hung Wong, 2001; Millenaar, Okyere, May, van Zanten, Voesenek, Peeters, 2006; Nielsen, Gautier, & Knudsen, 2005), attempt to normalize the data in such a manner in which the correlation between two samples becomes consistent, thereby allowing for easier identification of statistical outliers.

LOESS (Cleveland, 1979) is a local nonparameter method which attempts to fit a low order polynomial, normally linear or quadratic, to the scatter-plot attempting to minimize the random variations in the data. It is most often used for the *normalization* of two dye experiments, primarily to account for the slight difference in affinity between the two dyes at low expression levels, but can be used generally to correct for the nonlinearities found at lower intensity levels. It is similar to a nonlinear regression fit, except it performs a local regression upon blocks of data. The blocks of

Figure 1. The deviation from the log normal distribution at low intensity levels. The LOESS curve centers the distribution and forces log-normality (Wang, Hessner, Wu, Pati, & Ghosh, 2003)



overlapping data used are determined via a smoothing parameters which is set by the user under the constraints given in Equation 1**.** The size of the data block used for each local fit is then given as nq.

$$\frac{(d+1)}{n} \le q \le 1 \tag{1}$$

where d is the degree of the polynomial used.

Since LOESS is a local fitting method, the number of points used in each fit is nq, where q is the selected smoothing factor which satisfies the constraint in Equation 1 and n is the total number of points in the dataset. There is an additional weighting function given in Equation 2 which forces points closer to the point of estimation to contribute more to the local fit than those that are further away. The distance of the points to the point of estimation are calculated, and then normalized, so the data-point at the point of estimation has a value of 0, and the point furthest away has a value of 1. The primary advantage of loess is the fact that it is a non-parametric normalization and henceforth can be used when an *explicit model* of the data is not available.

$$w(x) = \begin{cases} \left(1-|x|^3\right)^3 & \text{for } |x| < 1 \\ 0 & \text{for } |x| \ge 1 \end{cases} \tag{2}$$

While the LOESS method is a nonparametric normalization of the data, dCHIP and PDNN (Position Dependent Nearest Neighbor), on the other hand, are model based approaches. In dCHIP (Li & Hung Wong, 2001), the primary normalization method creates a model of chip variability through the use of invariant marker genes, whilst PDNN (Nielsen et al., 2005) uses a nearest neighbor model to describe the physical binding of mRNA to the probes. The different relative binding of each probe to its exact match is then used to normalize the intensity value to obtain an expression level given the intensity. This method has the advantage of normalizing multiple probes for the same gene to roughly the same expression level. This is important because it was observed that oftentimes for time series, probes that had a different sequence that targeted the same gene had significantly different expression levels, despite being well correlated with each other. The problem with this issue is that oftentimes, one probe set will show statistically significant differential expression, whereas the other probe-set would not.

PDNN works by assuming that the observed signaling intensity comes from three sources, Binding, Non-Specific Binding, and background. It calculates the expected intensity via Equation 3

$$\bar{I}_{ij} = \frac{N_j}{1+e^{E_{ij}}} + \frac{N^*}{1+e^{E^*_{ij}}} + B \tag{3}$$

Where I is the expected signal intensity, i is the probe index, j is the gene index, N_j is the number of mRNA copies in the sample, N^* is the number of mRNA copies that contribute to Nonspecific Binding, E_{ij} is the free energy of formation for the specific binding and E^*_{ij} is the average non-specific free energy, and B is the background intensity.

E_{ij} and E^*_{ij} are calculated as the sum of the stacking energy for the binding which is calculated via the nearest neighbor model given in Equation 4.

$$E = \sum w_k \varepsilon(b_k, b_{k+1}). \tag{4}$$

The w_k represents the individual weights for each nearest neighbor pair given by ε. All of the parameters are computed by minimizing the following Equation 5.

$$F = \sum_i \sum_j \left(\ln \bar{I} - \ln I\right)^2 \tag{5}$$

The normalized expression for each gene then consists of N_j which would then be the predicted number of mRNA copies given the intensity values. This value of N_j is useful because it provides a method for reconciling probe sets which target the same gene, but due to variation in their mRNA sequence show different intensity profiles. This allows for the acquisition of a more accurate measure of mRNA activity than by just relying on intensity. Furthermore, genes represented by multiple probe-sets can have their predicted gene expression averaged to further reduce the contribution of noise in the overall signal, something which cannot be done when using intensity data only.

Comparisons to the different normalization methods (Millenaar et al., 2006; Ryden et al., 2006) have suggested that while they show some difference, it was found that the set of genes that were selected as differentially expressed show a significant level of concordance.

Selection of the Statistical Cutoff

In the selection of genes that show a statistically significant chance of being non-random, there is the question of p-values. Many researchers utilize a p value of $p < .05$, or $p < .01$(Lee, Kuo, Whitmore, & Sklar, 2000). However, it is important to consider the overall sample size. In a representative array, the Affymetrix RAE230A, there are over 15K probe spots on the array. With a p-value of .05, it still leads to the possible random selection of 750 genes, which represents a significant reduction in the number of genes being analyzed but may lead to results which would be difficult to justify statistically. Instead, the p-value ought to be set a p-value of $p < 1/N$ where N is the number of probes on the chip. For microarrays, we term this the "Natural" p-value since it directly relates the p-value to the sample size.

Gene Selection

The simplest filtering technique which researchers use to select for statistically significant genes is the 2-fold test. This test essentially selects for genes that have shown either a two fold up/down regulation, and is widely used because of its inherent simplicity. This test is essentially an approximation of the t-test with three samples at a confidence interval of $p < .05$. It tends to work poorly at low expression levels due to the greater variability vs. the mean, that is, a lower signal to noise ratio (Novak, Sladek, & Hudson, 2002) as well as the fact that the log normal distribution breaks down at the tail end corresponding to low expression levels. Therefore, before the use of the 2-fold selection method, there is usually a normalization step performed such as LOESS, as well as a filter for sample data quality to filter out the gene with a low expression level. However, despite the mathematical transformations to correct for the deviations from the log-normal distribution as well as the loss of the SNR, the 2-fold test represents a very weak filtering method because there is no explicit control as to what the confidence level of the selected genes are.

The problem with utilizing any sort of fold measure whether it be 2 fold, or n fold for greater or lesser stringency is that one is unable to report the confidence level of the selection because of the dependence of the confidence interval upon the variation between the different replicates. As a general rule, there are two competing trends with utilizing this selection:

1. The repeatability of the gene chips is increasing due to improvements in quality and technological improvements such as the use of 60 base pair recognition sequences(Hardiman, 2004), making the 2-fold test more stringent now than in the past.
2. The number of spots per array has increased, thereby increasing the natural p-value needed to separate out random changes in expression levels.

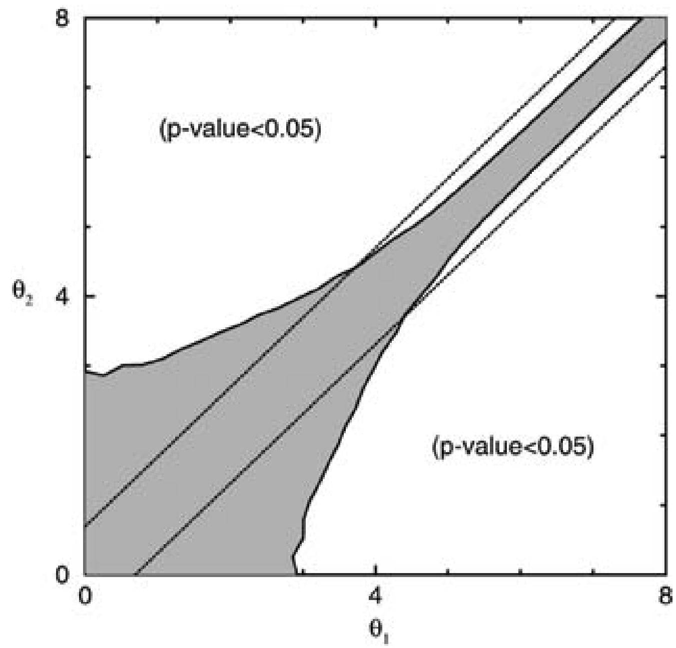
Both of these can be rectified with the use of the t-test which is give in Equation 6.

$$t = \frac{\overline{X_a} - \overline{X_b}}{\sqrt{\frac{\mathrm{var}(X_a)}{size(X_a)} + \frac{\mathrm{var}(X_b)}{size(X_b)}}} \tag{6}$$

The benefit of the utilizing the standard t-test is that the "Natural" p-value can be used as the cutoff for stringency as well as the fact that the metric is tailored for each individual gene with their attendant means and standard deviation. This method also has the added benefit of being able to handle the higher SNR found at the lower expression levels, Figure 2, because it does not assume the fact that there is a strict linear cutoff at all measured expression levels. Again as in the two-fold test, it is imperative that proper normalization take place before the use of the t-test, for without proper normalization, the any systematic errors that take place at lower expression levels tend to overwhelm the t-test.

A more sophisticated method for determining the quality of a gene's expression data is the Significance Analysis of Microarrays (SAM) method (Tusher, Tibshirani, & Chu, 2001). This method utilizes similar statistical metric to the t-test Equation 7. The primary difference between SAM and the standard t-test is the presence of an S_0 term which is calculated form the data that is designed to minimize the coefficient of variance of d(i) between the genes with low expression level and those with high expression levels. Additionally, instead of focusing on a statistical distribution such as the one

Figure 2. The variability between two chips. It is evident that as the expression level is lower, the linear trend between the two chips breaks down, and henceforth the 2-fold threshold does not become applicable. However, the t-test is able to correctly compensate for this change at lower expression level. Note however, that we do not advocate the use of the confidence interval of p < .05 for outlier detection(Tu, Stolovitzky, & Klein, 2002).



used for the t-test, the cutoff is determined arbitrarily by looking at a sorted plot of the d(i) values, Figure 3. Despite the fact that it is slightly different than that of the t-test, it satisfies the same goal which is primarily the assessment as to whether or not the difference measured between two samples has been reliably captured. The primary advantage of this method is that due to the normalization term S_0, there is increased sensitivity at lower expression levels, that is, it is less sensitive to SNR than the simple t-test.

$$d(i) = \frac{\overline{x_I}(i) - \overline{x}_u(i)}{\sqrt{a\left\{\sum_m [x_m(i) - \overline{x_I}(i)]^2 + \sum_n [x_n(i) - \overline{x_U}(i)]^2\right\}} + s_0} \quad (7)$$

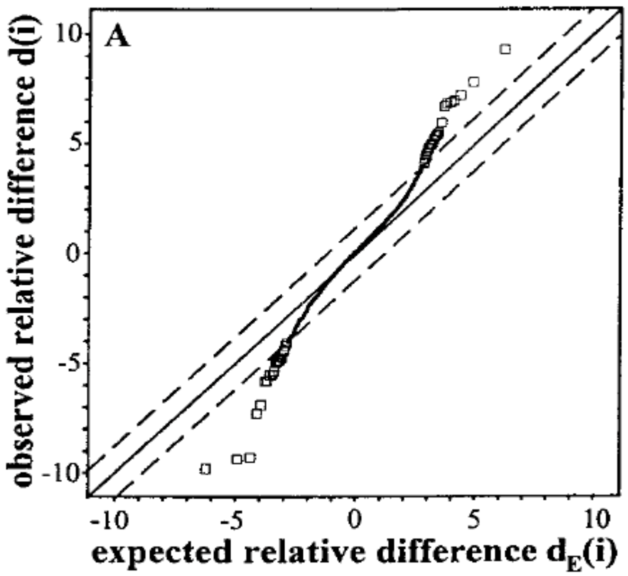
FUTURE TRENDS

One of the major pushes recently with *microarrays* is the use of temporal expression data. On a fundamental level, the same problem applies, namely the selection of statistically significant genes from the microarray. However, at this point it is not clear as to what a statistically significant temporal expression profile is. Some work has been done with the use of statistical over-representation of possible expression profile templates (Ernst & Bar-Joseph, 2006; Yang, Maguire, Yarmush, Berthiaume, & Androulakis, 2007) and future work has been proposed to make use of the replicates to identify those genes whose SNR in terms of *biological noise* is low enough where one can place high confidence in the overall shape of the obtained expression profile.

CONCLUSION

While microarrays have been hailed as a revolutionary advance in molecular biology, skeptics have pointed out the fact that there exists maddening inconsistencies in the results derived from microarray data (Kothapalli, Yoder, Mane, & Loughran, 2002). The primary problem with microarrays is the fact that at the same time, they are able to provide too much and not enough information. In the too much information realm, they provide the expression levels of thousands of genes at once, most of which are not related to the underlying phenomenon

Figure 3. The threshold's for SAM allows for the differentiation between genes that show differential expression as compared to the base state



being investigated. In the too few information realm, they are expensive enough where oftentimes there aren't sufficient replicates to deal with the inherent problem of technical and *biological noise*, making the selection of a set of genes questionable.

While technical noise is progressively being decreased through technological improvements via the makers of microarrays such as Affymetrix and Agilent, the problem of *biological noise* still remains. However, in either case, the ability to find experiments, genes, and probes whose intrinsic property remains despite the noise is still important. We feel that many of the criticisms of microarrays and the conclusions can be addressed with proper statistical evaluation of the data. This is not to say that conclusions that have not passed statistical muster do not contain important information; however statistical robustness adds another method of verification of the results. By utilizing experiment data quality assessments, one can determine whether an unknown process has indeed been captured in the context of both the biological response as well as technical issues such as the number of replicates, and the length of the time series. Proper analysis of the genes themselves will allow for the identification of genes that can be used for further analysis, and, finally, the analysis of the probes themselves ought to give an insight into the overall confidence in the values obtained from the technology as well as providing information as to whether or not the gene is actually expressed in vivo.

REFERENCES

Bowtell, D. D. (1999). Options available-from start to finish-for obtaining expression data by microarray. *Nat Genet, 21*(Suppl. 1), 25-32.

Brown, P. O., & Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nat Genet, 21*(Suppl. 1), 33-37.

Cheung, V. G., Morley, M., Aguilar, F., Massimi, A., Kucherlapati, R., & Childs, G. (1999). Making and reading microarrays. *Nat Genet, 21*(Suppl. 1), 15-19.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association, 74,* 829-836.

Dupuy, A., & Simon, R. M. (2007). Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J Natl Cancer Inst, 99*(2), 147-157.

Ernst, J., & Bar-Joseph, Z. (2006). STEM: A tool for the analysis of short time series gene expression data. *BMC Bioinformatics, 7,* 191.

Hardiman, G. (2004). Microarray platforms—Comparisons and contrasts. *Pharmacogenomics, 5*(5), 487-502.

Kothapalli, R., Yoder, S. J., Mane, S., & Loughran, T. P., Jr. (2002). Microarray results: How accurate are they? *BMC Bioinformatics, 3*, 22.

Lee, M. L., Kuo, F. C., Whitmore, G. A., & Sklar, J. (2000). Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proc Natl Acad Sci USA, 97*(18), 9834-9839.

Li, C., & Hung Wong, W. (2001). Model-based analysis of oligonucleotide arrays: Model validation, design issues and standard error application. *Genome Biol, 2*(8), RESEARCH0032.

Millenaar, F. F., Okyere, J., May, S. T., van Zanten, M., Voesenek, L. A., & Peeters, A. J. (2006). How to decide? Different methods of calculating gene expression from short oligonucleotide array data will give different results. *BMC Bioinformatics, 7*, 137.

Nakai, K., & Vert, J. P. (2002). Genome informatics for data-driven biology. *Genome Biol, 3*(4), REPORTS4010.

Nielsen, H. B., Gautier, L., & Knudsen, S. (2005). Implementation of a gene expression index calculation method based on the PDNN model. *Bioinformatics, 21*(5), 687-688.

Novak, J. P., Sladek, R., & Hudson, T. J. (2002). Characterization of variability in large-scale gene expression data: Implications for study design. *Genomics, 79*(1), 104-113.

Ryden, P., Andersson, H., Landfors, M., Naslund, L., Hartmanova, B., Noppa, L., et al. (2006). Evaluation of microarray data normalization procedures using spike-in experiments. *BMC Bioinformatics, 7*, 300.

Smyth, G. K., Yang, Y. H., & Speed, T. (2003). Statistical issues in cDNA microarray data analysis. *Methods Mol Biol, 224*, 111-136.

Tu, Y., Stolovitzky, G., & Klein, U. (2002). Quantitative noise analysis for gene expression microarray experiments. *Proc Natl Acad Sci USA, 99*(22), 14031-14036.

Tusher, V. G., Tibshirani, R., & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA, 98*(9), 5116-5121.

Wang, X., Hessner, M. J., Wu, Y., Pati, N., & Ghosh, S. (2003). Quantitative quality control in microarray experiments and the application in data filtering, normalization and false positive rate prediction. *Bioinformatics, 19*(11), 1341-1347.

Wu, J., & Androulakis, I. P. (2005). *Selecting maximally informative genes: The interplay between accuracy and complexity.* Paper presented at the 18th International Conference on Systems Engineering.

Yang, E., Maguire, T., Yarmush, M. L., Berthiaume, F., & Androulakis, I. P. (2007). Bioinformatics analysis of the early inflammatory response in a rat thermal injury model. *BMC Bioinformatics, 8*(1), 10.

Zhang, L., Miles, M. F., & Aldape, K. D. (2003). A model of molecular interactions on short oligonucleotide microarrays. *Nat Biotechnol, 21*(7), 818-821.

KEY TERMS

Biologically Informative: This is different from the notion of statistically significant because this set of genes is consistent over multiple experiments, replicates, and microarray platforms and reflects the underlying ground truth.

Locally Weighted Normalization of a Scatter Plot (LOESS; LOWESS): LOESS seeks to find a low order polynomial that best describes the overall variation in a scatter plot. This is used to normalize for the nonlinearities found in two state experiments.

Natural P-Value: The p-value a researcher should set in determining statistical significance. It is wholly reliant upon the number of samples in a given trial. Therefore, in a microarray, the natural p-value should be set to 1/N where N is the number of samples

Position Dependent Nearest Neighbor (PDNN) Model: A normalization technique by Zhang, Miles, and Aldape (2003), which makes the assumption that the signal intensity is dependent on both the probe sequence being used and the number of mRNA copies. It performs the normalization by calculating the number of mRNA copies and an expected signal intensity by optimizing for various parameters such as base stacking energy.

Significance Analysis of Microarrays (SAM): A selection algorithm which is nominally very similar to that of the t-test. It is, however, more robust to mRNA signals of lower SNR and hence gives more reliable filtering for genes of low expression levels.

Statistically Significant: The ability for the variability of a sample to be attributed by a factor other than through random noise. This is dependent first upon the overall distribution of the samples, though most researchers assume that the random variations are gaussian. Due to systematic factors such as dye binding affinities as well as the nonlinear binding behavior in microarrays, normalization is required before the use of this gaussian assumption

Signal to Noise Ratio (SNR): In the context of microarrays, the noise comes from two sources, technical and biological. This is, however, the primary determinant of how many replicates are required but is complicated via the fact that different probes have different SNR.

ATM Networks: Basic Ideas and Health

Petroudi Dimitra
University of Athens, Greece

Giannakakis Nikolaos
University of Athens, Greece

Poravas Efstratios
University of Athens, Greece

Robolas Periklis
University of Athens, Greece

INTRODUCTION

Telephone companies are finding themselves confronted with a very basic problem: multiple networks. The solution that was devised is a unique new network for the future that will replace the entire telephone system and all the specialized networks with a unique completed network for all types of transport of information. This new network will support enormous rhythm of data compared to all the existing networks and services, and will render possible the offer of a large variety of new services.

The new service network for a wide region is called Digital Network of Completed Services of Wide Area broadband integrated services digital network (B-ISDN). The technological infrastructure that renders it feasible is called asynchronous way of transport: asynchronous transfer mode (ATM) because the transmission is not modern; that is, attached in a being first clock.

GENERALLY FOR THE ATM

The ATM was materialized initially in order to be used in the WANs, however, its value quickly appeared for the LANs. The topology that is used is that of aster. The stations are connected via an ATM Switch. Each Switch can now be connected, with his line in some other hierarchically superior Switch that plays the role of backbone. It should be noted that these appliances allow the transfer of data of concrete breadth of area; however, capacities today reach 10Gbps.

The flexibility of ATM springs from the segmentation of parcels in smaller departments that are named cells. Each cell has a length of 53 bytes from the five constituting the heading (header).

The ATM Switch undertakes the promotion cells to the recipient after decoding the address of the destination from the heading of each cell.

BASIC IDEA

The basic idea behind the ATM is the transport of all information in small determined length parcels that are named cells.

The reasons for cell choice are many. Some of them selectively are:

- The cell transfer is exceptionally flexible and it can easily handle the movement of constant rhythm (e.g., sound, video) and the movement of variable rhythm (data).
- In the very high speeds that are expected, the digital cell transfer is easier than the use of traditional multiplex techniques.
- With regard to television distribution, the possibility of emission is essential. The cell transfer can support it, while the transfer of circuit cannot.

The ATM is based on protocols of virtual connection (e.g., virtual paths, virtual circuits). The user of the ATM has the possibility of accessing in entirety the breadth of area of channel whenever he or she wants for as long as he or she wants. The ATM ensures that services such as voice and moving pictures are transmitted in priority, while the longest time of waiting in the tail of transmission is the time of transmission of cell 53 bytes (roughly 3m sec in rhythm 155 Mbps).

The means of transmission for the ATM is usually the optical fibres, but for connections under 100 meters, coaxial cables or double conductor category 5 can also be used. The fibers can have a length of many kilometres. Each junction is extended between a computer and a transporter ATM or between two transporter ATMs. Each junction from point to point is one way. For completely bidirectional communication, they need two parallel junctions for the movement to each direction.

That is to say, the ATM Connections can become:

- Point to point (point-point)
- Point to a lot of points (point-multipoint)
- Points-to-points (multipoint-multipoint)

The forecasted speeds for ATM networks are 155 Mbps and 622 Mbps, with the possibility of speeds a Gigabit later. The speed of 155 Mbps became in order to be compatible with the system of transmission SONET. The speed of 622 Mbps was selected so with this can be sent four channels of 155 Mbps.

CATEGORIES OF SERVICES OF ATM

Category of constant rhythm bit constant bit rate (CBR) has as its aim to assimilate the cupreous cable or the optical fiber.

The next category of variable rhythm bit variable bit rate (VPR) is divided to subclasses for real time and not real time, respectively. Real-time variable bit rate (RT-VPR) is intended for services that have variable rhythm bits and is combined with strict requirements of real time, as the dialogic video. The other subclass, non-real-time (nrt-VBR) serves the movement; the convenient delivery is important, but a certain small quantity is bearable from the application.

Finally, we reach the unspecified bit rate (UBR), which does not promise anything and does not provide retroaction for the congestion.

USES OF ATM

The widespread deployment of high-speed networks has spurred the development of multimedia applications such as voice and video. In the medical domain, the transmission of medical images over networks opens up the possibility of improved education by allowing remote participation in clinical conferences, or improved and more cost-effective diagnoses by allowing remote consultations with experts. This development is aligned with trends such as the rise of managed health care organizations and the increased pressures for cost reduction in medical care. For example, the Hamad Hospital in Dubai (1998) has signed an agreement in order to organize the installation of a health care network based on asynchronous transfer mode (ATM) technology. This network would enable the hospital to take up a pioneering role in providing innovative telemedicine solutions throughout the whole of the region for many years to come. ATM can be used in:

- **LAN:** Network ATM can function as LAN that connects separate users or as bridges that connect more LAN.
- **WAN:** The ATM offers important faculties for the WAN networks:
    - efficient management
    - control of network
    - the ATM does not face problems with the distance
    - the integrity of a transmitted signal is even ensured when different types of movement are presented in the same network
    - it can offer different services in various speeds and a lot of levels of output
- **MAN:** The movement in a MAN network is limited in distances of certain kilometers.
- As network of trunk

The ATM had big success as a network. It supports a lot of different technologies such as:

- DSL
- IP Ethernet
- Frame Relay
- SONET

OTHER PARAMETERS THAT SHOULD BE EXAMINED

Extension of Network

The ATM:

a. It is capable of supporting services in a lot of different rhythms of transmission, and hence can be used in each part of a network.
b. It can transport each type of service in all the breadth of the network; after this, it is also the aim of his initial designing.
c. It has failed until now to prevail or be extended in an important degree in the desktop part of networks.

Cost

The ATM presents cost, which results because it is for a high quality technology that is used mainly by a limited crowd of institutions for high specifications and objectives networks.

Complexity

The ATM:

- Has complex and complicated products
- Has difficult installation
- Provides difficult management of network from the Ethernet
- Finds mistakes but also the comprehension connection—oriented ATM of networks very often constitutes a real headache

Quality of Service in the ATM

There are three types of ATM services: permanent virtual circuits, SVC, switched virtual circuits, and without connection. The quality of service is an important subject for network ATMs, partly because they are used for movement of real time as the sound and the video.

Safety in the ATM

An element that was not reported until now and constitutes an important factor of the reliability of the network is safety.

The ATM can and it precisely provides safety in the connections because the “circuit” that is installed with a connection is virtual and is decomposed immediately after the end of the connection.

Advantages of ATM

The ATM offers:

- A high control in the elements of transmission and their easy management
- Constant connections
- Guaranteed breadth of area
- Low and constant latency
- A high output via the transformation of material
- Evolution in the speed and the size of networks
- Support LAN/WAN technology
- International conformity of models
- Can be used in near or far distances

The services that the ATM network offers are:

- Video on demand
- Tile-education
- Telemedicine
- Virtual telephony
- Video conferencing
- Desktop conferencing
- Videophone
- Picture/sound at order (audio/video being demand)
- Virtual local networks (VLANs: Virtual LANs)
- Communications ATM of big capacity with mobile nodes (usually with satellite junctions)
- Interconnection of local networks
- Unification of voice and data (VoATM)

A

CONCLUSION

Numerous telcos have implemented wide-area ATM networks, and many ADSL implementations use ATM. However, ATM has failed to gain wide use as a LAN technology, and its great complexity has held back its full deployment as the single integrating network technology in the way that its inventors originally intended.

ADDITIONAL READING

http://ru6.cti.gr/bouras/dialekseis/3/atm2.ppt

http://www.syzefxis.gov.gr/gr/tech/atm.php

http://ru6.cti.gr/bouras/dialekseis/3/GIGABIT_2.ppt

http://www.lanl.gov/lanp/atm.tutorial.html#02

http://www.cse.ohio-state.edu/~jain/atm/index.html

http://www.atmforum.com/aboutatm/history.html

http://www.atmforum.com/aboutatm/guide.html

http://www.atmforum.com/

http://www.cne.gmu.edu/modules/atm/ATMstand.html

http://www.infosyssec.net/infosyssec/secatm1.htm

http://www.iec.org/online/tutorials/atm_fund/topic01.html

http://www.cisco.com/univercd/cc/td/doc/cisintwk/ito_doc/atm

http://www.npac.syr.edu/users/mahesh/homepage/atm_tutorial/

http://uranus.ee.auth.gr/report/gr/part2/chap9/9_5_5.html

http://www.techfest.com/networking/atm.htm

http://odysseia.cti.gr/e32-askoi/Enhmerwsh/Dihmerida/EisagwghATM/EisagwghATM_ie3.htm

http://www.medialab.ntua.gr/multi/atm/questions.html

http://www.medialab.ntua.gr/multi/atm/index.html#1

http://en.wikipedia.org/wiki/Asynchronous_Transfer_Mode

http://www.hoise.com/vmw/articles/LV-VM-03-98-10.html

KEY TERMS

Asynchronous Transfer Mode (ATM): A cell relay network protocol that encodes data traffic into small fixed-sized (53 byte; 48 bytes of data; and 5 bytes of header information) cells instead of variable sized *packets* (sometimes known as *frames*) as in packet-switched networks (e.g., Internet Protocol or Ethernet).

Computer Networking: The scientific and engineering discipline concerned with communication between computer systems.

Data T ransmission: The conveyance of any kind of information from one space to another.

LAN: Local area network.

Safety: The state of being safe, the condition of being protected against physical, social, spiritual, financial, political, emotional, occupational, psychological, or other types or consequences of failure, damage, error, accidents, harm, or any other event that could be considered undesirable.

Transmission (telecommunications): The act of transmitting electrical messages.

WAN: Wide area network.

Automatic Quantification of P-Wave Morphological Features

Federica Censi
Istituto Superiore di Sanità, Italy

Giovanni Calcagnini
Istituto Superiore di Sanità, Italy

Eugenio Mattei
Istituto Superiore di Sanità, Italy

Michele Triventi
Istituto Superiore di Sanità, Italy

Pietro Bartolini
Istituto Superiore di Sanità, Italy

INTRODUCTION

Atrial fibrillation (AF) is the most frequently occurring sustained cardiac rhythm disturbance (Wolf, Mitchell, Baker, Kannel, & D'Agostino, 1998). Although relatively easy to diagnose by analysis of the surface ECG, AF has modalities, mechanisms, and predisposing conditions that still remain poorly understood.

AF is not a direct life-threatening arrhythmia. However, because of the associated strong symptomatology, it frequently results in hospitalisation, physician visits, and drug therapy, hereby limiting the physical and social activities of many patients. It not only affects the quality of life, but it also increases the likelihood of prothrombotic effects and the risk for mortality because of cerebrovascular events or progressive ventricular dysfunction (Wolf et al., 1998).

AF is also a frequent post-cardiosurgery complication that results in an increase of length of hospital stay and of the associated costs (Kannel, Wolf, Benjamin, & Levy, 1998). Even though several underlying pathophysiological conditions might predispose to AF, no reliable method exists as yet to stratify the relative risk for AF development from a patient's clinical state, or by analysis of the available bedside data (Gang, Hnatkova, Mandal, Ghuran, & Malik, 2004). It is not possible to predict episode recurrence in paroxysmal patients, since AF can often be asymptomatic, nor does a method to identify patients at risk for post-cardiosurgery AF by pre-operative clinical investigations exist.

In the last decades, several studies have focused on finding reliable methods for the prediction of atrial fibrillation by analysis of surface electrocardiographic records. The ECG is a simple and widely available noninvasive technique used in the diagnosis of various cardiac diseases. The P-wave represents the sequential atrial activation of right and left atria, and it reflects the atrial conduction properties. Indeed, individuals with a clinical history of paroxysmal AF have a longer inter-atrial and intra-atrial conduction time, shown as a P-wave prolongation on the surface ECG. Moreover, the regional differences in atrial activation time in AF patients might be reflected in temporal variations across the leads of an orthogonal ECG (Villani, Piepoli, Rosi, & Capucci, 1996).

These findings have lead to an increasing interest in time domain P-wave analysis as a tool for AF risk stratification. Maximum P-wave (i.e., longest P-wave in a 12-lead ECG) and P-wave dispersion (difference between the longest and the shortest P-wave duration in 12-lead ECG) are currently the most sensitive indexes for AF risk assessment (Dilaveris & Gialafos, 2001). However, time domain methods are limited by the lack of a standard and accepted definition of P-wave duration. P-wave boundaries are difficult to detect, both because of the small amplitude of the atrial signal on the ECG, and because of the gradual slope of the atrial waveform from the baseline. For this reason, its determination is mainly attributed to the cardiologist's opinion and experience. Obviously, this reduces the applicability of time domain indexes in AF risk stratification, since cut-off values result to be dependent on the particular study group, the patient population investigated, and also the technological equipment used.

Nevertheless, irregularities of P-wave morphology have been detected in patients with paroxysmal atrial fibrillation (Dilaveris & Gialafos, 2002), and different shapes of P-wave may represent the presence or absence of an underlying pathophysiological condition in patients prone to AF attacks (Carlson, Johansson, & Olsson, 2001). Indeed, P-wave morphological analysis can help detecting interatrial blocks, which predispose to this cardiac rhythm disturbance, even in patients who develop AF despite the absence of particular alterations in echocardiographic parameters (Bayes de Luna, Guindo, Vinolas, Martinez-Rubio, Oter, & Bayes-Genis, 1999).

BACKGROUND

Morphological analysis of P-wave has been extensively used to assess interatrial conduction defects and abnormalities of the left and right atrium (Dilaveris & Gialafos, 2002; Michelucci, Bagliani, Colella, Pieragnoli, Porciani, Gensini, Padeletti, 2002; Platonov, Carlson, Ingemansson, Roijer, Hansson, Chireikin, & Olsson, 2000). Changes in P-wave polarity, as well as subtle differences in P-wave morphology, are believed to reflect abnormal activation patterns in the atria (Michelucci et al., 2002). Irregularities in the orthogonal P-wave morphology have been detected in AF patients and associated with local interatrial conduction delay (Platonov et al., 2000). Different shapes of P-wave have been shown to represent the presence or absence of an underlying pathophysiological condition in patients prone to AF attacks (Gang et al., 2004). In the presence of a partial inter-atrial conduction block, P-wave shape changes have been demonstrated to depend on the direction of the activation wavefront. The studies on P-wave morphology are usually performed by visual inspection of shape changes. However, abnormalities of left or right atrium are likely reflected in subtle morphological changes beyond visual classification.

We hereby present a P-wave model, based on a linear combination of Gaussian functions, to perform an automatic and reliable quantification of the morphological features of P-wave in patients with paroxysmal atrial fibrillation (Censi et al., 2007). In addition, given the importance of multichannel electrocardiography for the extraction of quantitative parameters, we used a 32-lead mapping ECG system, to increase the spatial sampling on the body surface and to enhance low amplitude signals (Trobec, 2003).

MORPHOLOGICAL ANALYSIS OF THE P-WAVE

In order to overcome the limitations related to the time-domain analysis, an automated morphological analysis of the P-wave can be obtained by a model based on a Gaussian fit of averaged P-wave. Such a model can numerically and automatically quantify, beyond visual inspection, morphological aspects of P- wave in AF patients.

By describing a P-wave as a linear combination of up to eight Gaussian functions, the number of Gaussians that explain 97.5% of the signal variance is used as representative of that specific wave. The fit, in a way, acts as a filter that eliminates high-frequency noise.

Since more than one Gaussian function can be used by the model to cover steep portions of the signal, without really fitting a peak or valley, a better morphological description of the wave is obtained by evaluating the number of *maxima, minima,* and *zeroes* of the extracted fit. The P-wave morphology is then evaluated by the following parameters: the model order at which AdjRsq≥97.5%; the minimum (σ_{min}) standard deviation of the Gaussians included in the model; the maximum (σ_{max}) standard deviation of the Gaussians included in the model; the number of relative maxima and minima (max+min); and the number of zero crossings of the model (zeros). The Gaussian fitting of two P-waves with different morphology is shown in Figure 2. The P-wave on the left achieves an AdjRsq = 98.93% with a second first order model (two Gaussian functions), while in the second case, AdjRsq higher than 97.5% is only reached with a 6th order Gaussian fit. Lower panel shows the values of the five parameters extracted from the Gaussian fit to quantify the morphological features.

The parameters chosen synthesise the main morphological characteristics of the original waveforms: the order of the model is the number of Gaussians required; the number of maxima and minima provides information on the number of 'bumps' included in the waveform; the number of zeros accounts for the phase changes of the P-wave.

Figure 1. Electrode positioning scheme

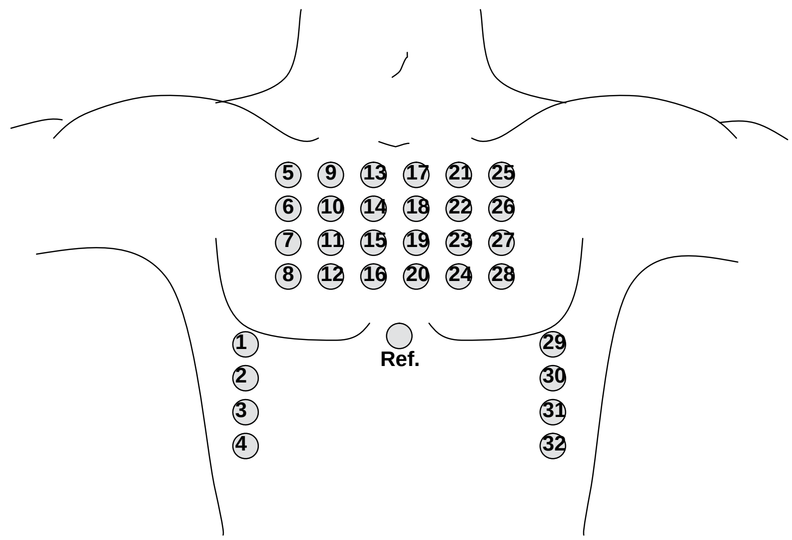
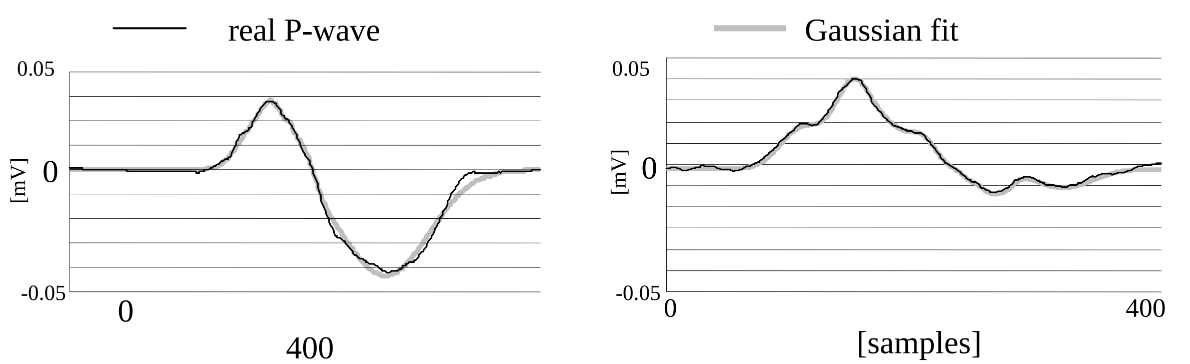


Figure 2. Gaussian fit of two P-waves with different morphology. Upper panel – left: two Gaussian functions fit the P-wave with AdjRsq =98.93%. Upper panel – right: AdjRsq higher than 97.5% is only reached by a 6th order model. Lower panel: values of the five parameters extracted from the Gaussian fit to quantify the morphological features.



Best order: 2	*Best order: 6*
Zeros: 1	*Zeros: 1*
Max+min:2	*Max+min:4*
σ_{max}: *53.53*	σ_{max}: *53.78*
σ_{min}: *30.69*	σ_{min}: *23.20*

Table 1. Results for the morphological analysis expressed as mean values ± standard deviations and p-value obtained according to the statistical analysis explained in the text

	N	LR	HR	population *p*-value	N vs. LR *p*-value	N vs. HR *p*-value	LR vs. HR *p*-value
best model order	2.7±0.2	5.7±0.7	6.6±0.7	<0.01	<0.01	<0.01	<0.01
zeroes	0.7±0.3	0.9±0.34	1.8±0.7	<0.05	<0.05	<0.01	<0.05
max+min	3.1±0.4	6.1±1.01	7.3±1.5	<0.01	<0.05	<0.01	<0.05
sigma_max	46.8±6.7	50.8±6.2	51.0±10.2	0.4	0.2	0.2	0.9
sigma_min	24.5±2.2	22.2±10.7	23.5±9.6	0.5	0.2	0.6	0.5

ECG signals were acquired using a 32-lead mapping system for high-resolution biopotential measurement (ActiveTwo, Biosemi, the Netherlands), sample frequency 2 kHz, 24 bit resolution, 0-400 Hz bandwidth. The system is made of a battery powered AD box that digitises the signals and transfers them to a PCI receiver on computer through fiber optic connection. The 32 leads guarantees spatial sampling on the body surface. The leads were strategically positioned as shown in Figure 1, to allow accurate recordings of atrial signals. ECG recordings were acquired as single-ended signals, with respect to the common reference.

Twenty-five patients with permanent dual chamber pacemakers (AT500-Medtronic, Inc. Minneapolis, MN, USA) were recruited at South Filippo Neri Hospital, Rome, Italy. The AT500 device combines the monitoring of atrial tachyarrhythmias with both intervention therapies and pace-termination therapies. Intervention therapies aim at suppressing atrial tachyarrhythmia trigger mechanisms by using three programmable pacing algorithms. In addition, three pace-termination algorithms recognize treatable atrial tachycardias and delivers antitachycardia pace-therapies to restore sinus rhythm.

The study population consisted of 12 females and 13 males, aged 73 ± 8, affected by paroxysmal atrial fibrillation and bradycardia. Four healthy subjects, with no history or sign of cardiac disease, were recorded as a control group. Five-minute ECG recordings were performed for each subject, with the pacemaker programmed in VVI mode (i.e., in single-chamber ventricular pacing mode), set at 40/min. in order to have spontaneous beats. To compare the morphological parameters extracted from different type of patients, three classes of risk have been defined, according to the number of AF episodes recorded by dual chamber pacemakers in the last six months preceding the acquisition: class N: normal, no episodes and no risk associated (i.e., control subjects); class LR (low risk): number of episodes=0; class HR (high risk): number of episodes≥1.

Every lead signal was preprocessed and analysed to extract the average P-wave characteristic of the specific lead. The first step is to detect P-waves from the acquired signals. P-waves are isolated, considering 200ms windows triggered from the R wave of the QRS complexes. R-waves are detected using an algorithm similar to that proposed by Pan and Tompkins, which actually acts as a high pass filter on the ECG signals, enhancing the high frequency QRS complexes (Pan & Tompkins, 1998). Secondly, a beat-by-beat linear piecewise interpolation was used to remove baseline wander on each P-wave. Then, a coherent averaging procedure has been implemented. In order to take into account the variations in PR interval and/or the inaccuracy in R-wave detection, care had been paid to align the P-waves before the averaging procedure; particularly, P-waves were aligned according to the lag at which the cross-correlation function between the current averaged P-wave, and each single P-wave shows its maximum.

Mean values of these parameters over the 32 leads have been considered (Table 1). P-waves from the control group can be described by lower order models (2.7±0.2), compared to both LR (5.7±0.7) and HR patients (6.6±0.7). Also, morphological differences between LR and HR, evaluated in terms of model order, resulted to be significant ($p<0.01$). The number of relative maxima and minima from the Gaussian fit is significantly higher in HR class (7.3±1.5), compared to both LR (6.1±1.0) and N (3.1±0.4). The number of zeroes is also significantly increased in HR, compared to LR. No particular differences have been observed for σ_{max} and σ_{min} parameters.

These parameters might all be markers of the fractionated electrical activity that characterises paroxysmal AF patients in sinus rhythm. P-waves in control group, as expected from visual inspection, appear smoother than those from AF patients.

To our knowledge, no previous studies have performed a similar analysis in paroxysmal AF patients, thus no comparisons are possible. Gang and colleagues have recently proposed some complex morphological descriptors of the P-wave for prediction of AF risk in post-CABG patients, based on singular value decomposition of the ECG signal (Gang et al., 2004). These descriptors take into account global shape abnormalities and dissimilarities between shapes in individual leads. These parameters increased in AF patients, showing morphological alterations in this group, but failed to independently predict postoperative AF. A polynomial representation of P-wave was implemented by Clavier and colleagues, who developed a method for automatic analysis of the P-wave based on the use of a hidden Markov model and wavelets transformation (Clavier, Boucher, Lepage, Blanc, & Cornily, 2002).

A polynomial interpolation can only partially represent details in P-wave morphology that are thought to reflect the atrial conduction abnormalities, and degrees higher than the fourth generate waveforms that do not have a limited temporal extension, a characteristic that is indeed preserved by a linear combination of Gaussians. In addition, the Gaussian functions-based model of P-wave proposed in this study has the great advantage of being independent on the definition of P-wave boundaries, thus it can provide useful information in AF risk assessment.

FUTURE TRENDS

A fully automated analysis and quantification of P-wave morphological features could be useful for AF risk stratification, and could improve the sensibility and specificity of screening methods based only on the time-domain approach. Further studies are necessary to confirm the predictive value of the proposed descriptors of P-wave morphological characteristics using bigger population including control patients and prospective follow up of patients with different number and duration of AF episodes.

Electrocardiographic potentials recorded from multiple sites on the torso seem to help in yielding more diagnostic content by extracting information from anatomical regions not interrogated by the most common 12-lead ECG (Yamada, Fukunami, Shimonagata, Kumagai, Sanada, Ogita, Asano, Hori, & Hoki, 1999). The problem of optimal electrode positioning has not an unique solution, since it may depend on the kind of analysis to be performed (e.g., surface map, inverse problem) as well as by the pathology/population to be investigated, as reported by Finlay et al. (2006). However, there is general agreement and there are evidences that more electrodes should be placed on the anterior surface around the precordial area close to the heart. Given our necessity to record the electrical activity of the atria, we decided to place 24 out of 32 electrodes in the upper part of the torso, more concentrated in the area close to the heart; we also provided leads from the lateral zones by placing electrode A1-A4 and A29-A32. Whether this configuration is the optimal lead set for our population was not an objective goes beyond the aim of this study, and will be the objective of future investigations. We found that in many patients, the morphological parameters obtained from vertically aligned electrodes (e.g., A1-A4, A5-A8, and so on..., see figure 1) were rather similar. Interestingly, we found that the most discriminant leads for the morphological parameters are those placed on the right of the torso. Specifically designed configurations could thus be tested, with electrodes distributed also in the lower part of the torso and on the posterior surface, in order to investigate how the morphological parameters depend on the number and location of the leads.

Whether similar results can be obtained using conventional 12-lead ECG or Frank leads may be of clinical interest. Although the method itself can be applied to any lead configuration, our results were obtained using a wide-band, DC coupled, 24 bit resolution acquisition system with active electrodes. The impact of these technical features on the extracted parameters needs a deeper investigation.

CONCLUSION

Standard approaches for P-wave analysis deal mainly with the analysis of the duration of the P-wave. However, there is the lack of good and accepted definitions of P-wave onset and offset, whose determination cannot be automatic, and has a high interobserved variability.

Indeed P-wave morphology can be effectively modeled by a linear combination of Gaussian functions (i.e., by a Gaussian fitting). In addition, the morphological parameters extracted from the Gaussian function-based model of the P-wave can identify patients having low and high risk of developing AF.

REFERENCES

Bayes de Luna, A., Guindo, J., Vinolas, X., Martinez-Rubio, A., Oter, R., & Bayes-Genis, A. (1999). Third-degree inter-atrial block and supraventricular tachyarrhythmias. *Europace, 1*(1), 43–46.

Carlson, J., Johansson, R., & Olsson, S. B. (2001). Classification of electrocardiographic p-wave morphology. *IEEE Trans Biomed Eng, 48*, 401–405.

Dilaveris, P. E., & Gialafos, J. E. (2001). P-wave dispersion: A novel predictor of paroxysmal atrial fibrillation. *A.N.E., 6*, 159–165.

Dilaveris, P. E., & Gialafos, J. E. (2002). Future concept in p-wave morphological analyses. *Card Electrophysiol Rev, 6*, 221–224.

Dilaveris, P. E., Gialafos, J.E., & Sideris, S. (1998). Simple electrocardiographic markers for the prediction of paroxysmal idiopathic atrial fibrillation. *Am Heart J, 135*, 73–78.

Finlay, D. D., Nugent, C. D., Donnelly, M. P., Lux, R. L., McCullagh, P. J., & Black, N. D. (2006). Selection of optimal recording sites for limited lead body surface potential mapping: A sequential selection based approach. *BMC Med Inform Decis Mak., 17*, 6–9.

Gang, Y., Hnatkova, K., Mandal, K., Ghuran, A., & Malik, M. (2004). Preoperative electrocardiographic risk assessment of atrial fibrillation after coronary artery bypass grafting. *J Cardiovasc Electrophysiol., 15*, 1379–1386.

Kannel, W. B., Wolf, P. A., Benjamin, E. J., & Levy, D. (1998). Prevalence, incidence, prognosis, and predisposing conditions for atrial fibrillation: Population-based estimates. *Am J Cardiol., 16, 82*(8), 2–9.

Clavier, L., Boucher, J. M., Lepage, R., Blanc, J. J., & Cornily, J. C. (2002). Automatic p-wave analysis of patients prone to atrial fibrillation. *Med Biol Eng Comput., 40*(1), 63–71.

Michelucci, A., Bagliani, G., Colella, A., Pieragnoli, P., Porciani, M. C., Gensini, G., & Padeletti, L. (2002). P-wave assessment: State of the art update. *Card Electrophysiol Rev., 6*, 215–220.

Pan, J., Tompkins, W. J. (1998). A real-time QRS detection algorithm. *IEEE Trans Biomed Eng., 32*, 230–236.

Platonov, P. G., Carlson, J., Ingemansson, M. P., Roijer, A., Hansson, A., Chireikin, LV., & Olsson, S. B. (2000). Detection of inter-atrial conduction defects with unfiltered signal-averaged p-wave ECG in patients with lone atrial fibrillation. *Europace, 2*, 32–41.

Trobec, R. (2003). Computer analysis of multichannel ECG. *Comput Biol Med., 33*(3), 215–26.

Villani, G. Q., Piepoli, M., Rosi, A., & Capucci, A. (1996). P-wave dispersion index: A marker of patients with paroxysmal atrial fibrillation. *Int J Cardiol., 26, 55*(2), 169–175.

Wolf, P. A., Mitchell, J. B., Baker, C. S., Kannel, W. B., & D'Agostino, R. B. (1998). Impact of atrial fibrillation on mortality, stroke, and medical costs. *Arch Intern Med., 9, 158*(3), 229–234.

Yamada, T., Fukunami, M., Shimonagata, T., Kumagai, K., Sanada, S., Ogita, H., Asano, Y., Hori, M., & Hoki, N. (1999). Dispersion of signal-averaged p-wave duration on precordial body surface in patients with paroxysmal atrial fibrillation. *Eur Heart J., 20*(3), 211–220.

KEY TERMS

Atrial Fibrillation: Cardiac arrhythmia involving the atria, which are exited by irregular disorganized electrical impulses, causing atria not to regularly contract.

Automatic Quantification: Extraction of numerical information by an algorithm, without the intervention of an operator.

Gaussian Fit: Fitting procedure using Gaussian functions.

Morphological Quantification: Extracting numerical information associated to the morphology of the signal examined.

Multisite ECG Mapping: Electrocardiographic acquisition from multiple sites.

P-Wave: ECG wave associated to the atrial depolarization.

P-wave Abnormalities: Irregularities on the P-wave normal morphology.

Balancing the Capacity in Health Care

Lars Odd Petersen
Danish Institute for Health Services Research, Denmark

INTRODUCTION

Diagnostics and treatment of patients often involve several different clinics. When improving quality and efficiency in health care we therefore need to consider the entire pathway from the patient's first contact with a health institution until the final discharge from the hospital and rehabilitation after the end of treatment. A system view is needed to consider the stochastic nature of acute patient arrivals, the variety of pathways through the clinics, and the stochastic time needed for diagnostics and treatment.

A balancing of the capacity at the different clinics, in order to deliver health care services in due time, to improve the overall productivity or use of capacity, calls for a reallocation of resources. A successful accomplishment of such a process requires a common insight and agreement by the clinics and therefore representatives from the clinics participating in the analytical process. The clinician participation enables elucidation and use of information on organisational behaviour of importance for the daily operations. This type of information does not appear from the registered patient data and implies analytical tools, which are both simple and intuitive and capable of handling and displaying the type and amount of information needed in a trustworthy and consistent way.

BACKGROUND

Different concepts for organisational improvements developed in an industrial environment have been introduced to—and applied in—the health care sector over the years. We shall briefly comment on four of them.

In Business Process Reengineering (Hammer & Champy, 1993), a fundamental rethinking and radical redesign of business processes is in focus. The word processes is the keyword. It is also the word that gives corporate managers the greatest difficulties. Most business people are not "process-oriented." They are focused on tasks, on jobs, on people, on structures, but not on processes. A business process is defined as a collection of activities which takes one or more kinds of inputs and creates an output of value to the customer.

An indispensable tool in Business Process Reengineering is process mapping (Johanneson, McHugh, Pendlebury, & Wheeler, 1993). Competitive realignment through identifying and exploiting "Break Points" is achieved by reengineering core business processes. This, in turn, requires an extensive understanding of the activities which constitute the core business processes and the processes which support them, in terms of their purpose, trigger points, inputs and outputs, and constraining influences. This understanding can be achieved by "mapping," "modelling," and then measuring the processes using various techniques that have been developed and refined over the years from simple flow diagrams to advanced simulation modelling charts.

With total quality management (TQM), inspired by Deming's "Out of the Crisis" (Deming, 1982) followed a systematic work describing goals for quality, preventing adverse events, monitoring quality, and constant improvements in quality. In health care many efforts to improve the quality are based on developing care pathways (deLuc, 2001), which specify how the patients' flow through the health organisation should be in order to achieve good quality. From mapping the pathways for the various types of patients, good quality is specified, actually delivered quality is monitored and ways to improve the quality are constantly considered. Mapping and diagramming are important ingredients here too. In TQM, the system's view is in focus as well. It is a mistake to assume that if everybody does his job, it shall be all right. The whole system may be in trouble. In much of the TQM activity in health care, the aim is to specify good quality in standard pathways. However, the capacity needed to deliver the desired quality is very seldom brought into focus.

The concept of the Learning Organisation (Senge, 1990) is to a considerable extent based on systems thinking and mapping tools such as the "Causal Loop

Diagram (CLD)" and the "Stock and Flow Diagram" (SFD). Both types of mapping are core elements in the modelling theory called System Dynamics (Sterman, 2000). According to the ideas of the Learning Organisation, the diagrams play an important role in involving the managers and employees in the efforts to improve the organisation. Its abilities to elicit peoples' mental models of how things are run in their part of the organisation, to share information and different views and to clarify necessary conditions for organisational improvements is stressed. With the CLDs and SFDs it is possible to pick up most of the soft and hard data (information) about how the organisation works. In relation to the SFDs, several simulation programmes have been developed to handle and apply the organisation's knowledge consistently in simulating the consequences of organisational changes before their implementation.

In Lean Thinking (Womack, Jones, & Roos, 1990), there is a focus on the value stream, for example, the flow of patients from the contact before diagnosing to the end of treatment. If the work in the organisation does not create value for the patients, it is regarded as waste and should be eliminated or at least reduced to a minimum. The same goes for obstacles preventing a continuous flow of patients in the process of diagnosing and treating.

Despite the differences explained in this short and not necessarily exhaustive list of concepts for organisational improvements, they have much in common when it comes to the mapping tools and techniques. When dealing with changes in health care organisations, I agree with those (Hsieh, Tan, & Lin, 2003; Reidd, Compton, Grossmann, & Fanjiang, 2005) who stress the importance of using a system approach and selecting tools which enable the analyst to work with a system rather than independent entities, that is, departments and processes.

The need in health care to do analysis from a systems perspective, implying the use of hard, soft, and stochastic data, in a coherent and consistent process of analysis and the involvement of clinicians in the modelling process, lead us to focus on the systems thinking (system dynamic) tools.

One can find recommendations from clinical professionals for using systems thinking as a tool for development in the health care sector in general (Beirema 2003; Nolan 1998), and system dynamic is applied in various health care contexts such as accident and emergency departments (Lane, Monefeldt, & Lund, 2000), queuing problems (Waaler & Iversen, 1998), waiting lists (Gonzales-Busto & Garcia, 1999; Van Ackere & Smith, 1999), and in the development of new health care initiatives (Homer, Hirsch, Minniti, & Pierson, 2004; Wolstenholme, 1999). The documentation of the applicability of system dynamics in health care is increasing. In a resent publication from the U.S. Academy of Science System Dynamics is on the list of systems engineering tools recommended for use in the process of improving the health care sector in the country (Reidd et al., 2005).

THEORY AND METHODOLOGY

As a modelling approach system dynamics has three characteristics (Lane, 2000). First is the use of "feedback loops." They are present in every system and often identified and visualised in simple causal loop diagrams. In health care, it is often reasonable to let these diagrams show the flow of patients and the variables or information controlling the flow—such as the capacity or the utilisation of capacity. In general terms, they show how information is followed by controlling action based on policies, followed by a new information collection. These closed loops of causal links involve delays and nonlinear relationships.

The second characteristic is computer simulation. Although humans can conceptualize system structure and loops, they lack the cognitive capability to deduce the consequent dynamic behaviour of the system without assistance. Computer simulation is therefore used rigorously to deduce the behavioural consequences over time of the hypothesized causal network (flow of patients). Such behaviour sometimes turns up to be counterintuitive, and may be explored further using simulations models.

The third and last characteristic is the need to engage in mental models. The most important information about social institutions is only held as "mental models," and is not written down. These mental models are complex and subtle, involving hard, quantitative information, and more subjective or judgemental aspects of a given situation. Such models are the basis for organisational decision-making. Hence eliciting, debating, and facilitating change in the mental models of decision-makers can result in improved ways of managing a system. Modelling work must therefore be carried out in close proximity to problem owners, who are then able to see that their mental models are reflected in a computer model.

A distinguishing feature of system dynamics is the use of various methods for mapping to conceptualise, represent, and communicate the underlying models. Although there is a range of mapping approaches in use in the world of system dynamics, two methods are overwhelmingly accepted by the international system dynamics community. First, broad representations of variables and feedback structure of a model are conveyed using "causal loop diagrams" (CLD). In contrast, "stock/flow diagrams" (SFD) are more detailed, discriminating both state and flow variables.

"System dynamics" uses models to explore the link between system structure and time evolutionary behaviour. The aim is to elicit and develop the mental model of the system in focus, to explain system behaviour by providing a causal theory, and to use that theory as the basis for interventions into the system structure which will then change the resulting behaviour mode.

In Figure 1, an example of CLD is shown. An increase in the arrival of planned and acute patients makes the number of beds occupied increase ("+"). With a certain number of beds available, the use of beds will increase too and when beds used reaches the number of beds usually available, extra beds from other wards are asked for. In case of extra beds available, the total bed capacity is increased and the use of bed capacity decreases ("-"). If no extra beds are available, there is an opportunity either to decrease the number of planned patients called in for treatment or to force the discharge of patients, that is, by decreasing the average length of stay.

Figure 1 represents a typical example of the use of hard and soft data. The soft—but still quantitative part—is the specification of how they react in the bed ward when the use of capacity is above an acceptable level. In Figure 1, three loops are marked with the symbol "(-)" expressing that the loops are stabilising or balancing. Here it means that when the bed ward is under pressure (a high use of bed capacity), they react in one or several ways to compensate for that pressure.

In system dynamics there are two types of causal links ("+", "-") and three types of variables. Some variables are changed only by "flows" accumulating into or draining out of them. Such variables are the "state variables" of the model (called "stocks" or "levels"). In contrast, the quantities of the flows (called "rates" or "flows") are instantaneously causally established by "information links" from "ordinary variables" and other state variables.

Figure 2 shows a stock and flow diagram of the system mapped in Figure 1. The dynamic simulation software package I think is applied in making the diagram. Planned and acute patients flow into the department treating the patients. After staying there for a while—determined by the actual length of stay—the patients are discharged. The length of stay can change when the use of capacity is too high. The other two options for system stabilising reaction are also shown in Figure 2.

Beds occupied and the total bed capacity are both represented by a stock symbol (rectangular). The flow

Figure 1. Example of causal loop diagram (CLD)

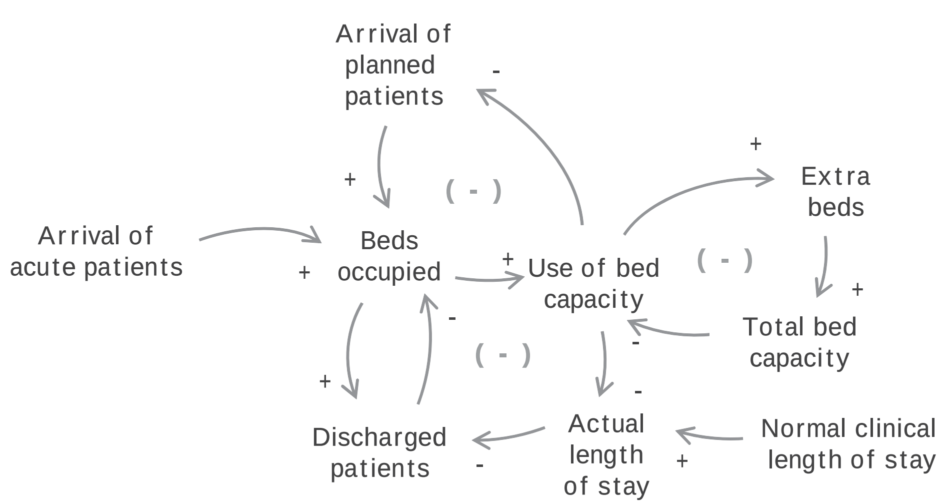
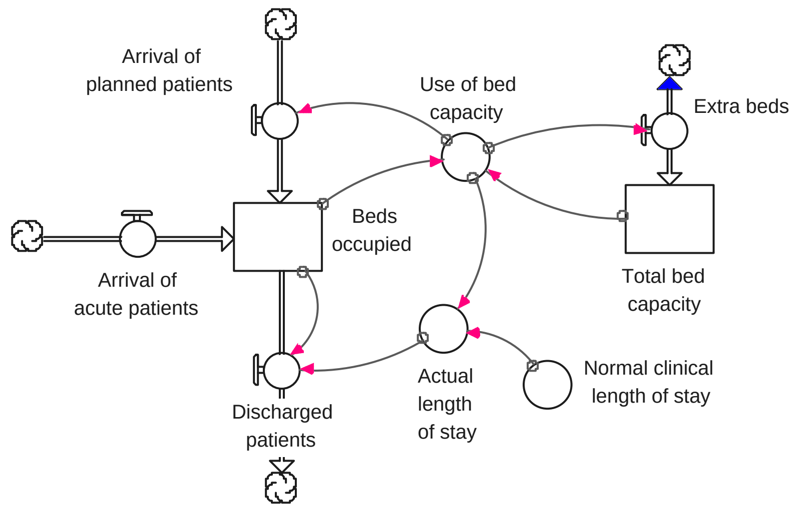


Figure 2. Example of stock and flow diagram (SFD)



symbol (valve) is applied to represent the flow of patients in and out of the bed ward. Extra beds are also shown as a flow variable that either increase or decrease the total capacity. The other (auxiliary) variables shown in Figure 2 are represented by a circle and hold the logic or information needed to make the system of stocks and flow running.

The advantages of CLDs are the very limited library of symbols and the concentration on loop structure. As conceptualisation tools they are simple, easy and enjoyable for nontechnical people to use.

The SFDs offer a much sounder basis than CLDs for the rigorous deduction of dynamic behaviour.

In this section, it is argued for the choice of System Dynamics as the methodological basis in the analysis of the capacity in health care organizations. The System Dynamic tools—CLD and SFD—are introduced and their applicability in a health care context is illustrated.

In the following section, we describe how System Dynamics tools are applied in a case. Afterwards, the use of SD in the analysis and in health care analysis in general is evaluated and discussed.

CASE

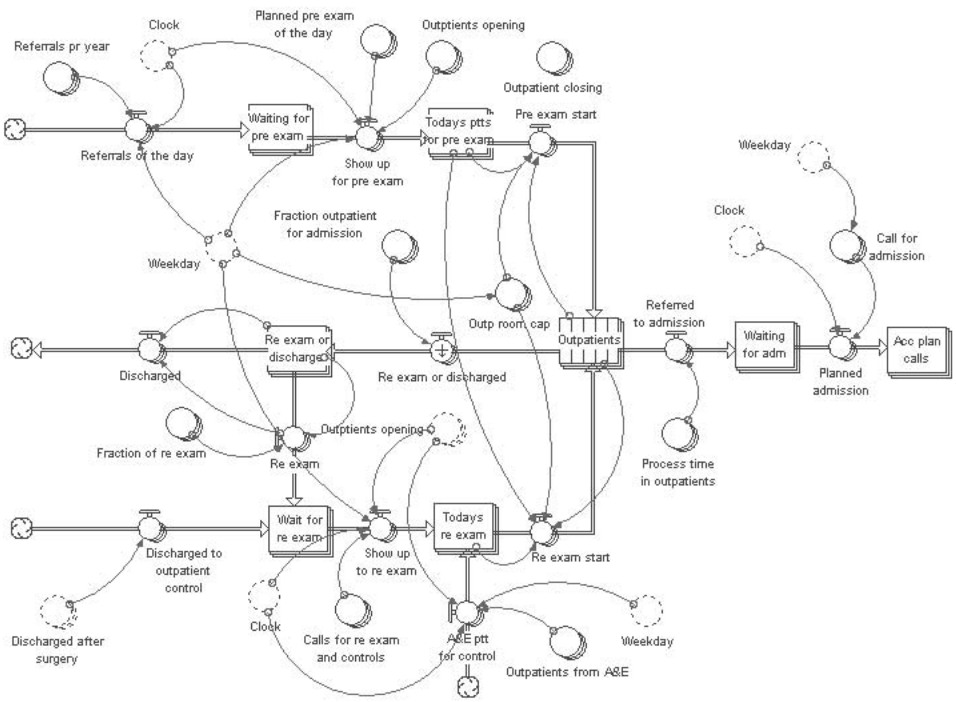
The mapping presented above was applied at a local hospital in the Copenhagen Metropolitan Area in an analysis of the surgical capacity with the intention to find out the extent to which the daily operations in the surgical departments could be improved. A benchmarking process had indicated a potential for improvements. The hospital had two surgical specialities—orthopaedics (O) and gastroenterology (G)—both sharing the same surgical department with four theatres (operating rooms) and a recovery section with four beds.

The analysis was carried out in collaboration with representatives from the surgical departments, anaesthetics, the operation theatres, and administration and assisted by an external consultant. Together they formed a working group. A small group of secretaries prepared the information for clarification in the working group and other specific tasks during the process such as data processing, mapping and computer modelling.

At a number of meetings in the working group during the autumn of 2004 the mapping of the surgical operations was done. This included drawing diagrams, going through quantitative data describing the operational details and deciphering the qualitative information about the behaviour in the system. During the process the daily operations were evaluated, and through the use of a simulation model, proposals for changes were described.

Figure 3 is a visual representation of one part of the model—the outpatient clinic—built up by using stock and flow symbols. Wards, Operating Theatres, Recovery and Accident, and Emergency are modelled in the same way but not shown in Figure 3. The stocks represent the number of patients in a certain stage in the patient flow—for example the number of patients waiting to be admitted, the number of patients in a ward or the

Figure 3. Visualisation of the outpatient clinic part of the simulation model



accumulated number of patients admitted. The large arrows (flows) represent the flow of patients going from one stage to another—for example, patients referred to the waiting list for admission. The circles represent variables in the model influencing the flow or logical relationships between internal variables such as stocks and flows and external parameters. Dashed symbols are copies of other variables. These are useful when one wishes to split the model into sections.

The flow of planned patients is seen as incoming referrals. All of them are called in for a pre-examination when capacity in the outpatient clinic allows it.

At first the model is used to calculate the actual figures describing the daily operation at the surgical departments. Comparing the recorded numbers with the simulated data of the actual operation showed a satisfying similarity.

From going through the operational details, numbers, and from modelling the actual daily operations, it was concluded that it should be possible to obtain a better balance in the use of capacity through

1. An improved flow of patients in theatres
2. A more flexible allocation of theatre capacity between specialities
3. A reduced length of stay on recovery
4. More beds in recovery
5. Accelerating the patient flows in department O by implementing Joint Care

From these initiatives, the model showed that it should be possible to increase the number of planned admissions of O patients by 36% and reduce the number of periods in recovery with more patients than allocated beds by approximately 25%.

FUTURE TRENDS

Concepts for organizational development emerge and many almost vanish after some years. Though the changes follow the fashion of management in appearance, some of the interior parts in shape of tools for mapping and analysing seem to recur in new versions often due to the PC development.

For decades examples has shown how health care can benefit from using operational management tools such as queuing theory and simulations in determining the level of capacity, improving the use of capacity, and many other managerial challenges (Bolling, 1972;

Schmitz & Kwak, 1972). Years of experiences tend to point to simulation as the tool that best can handle models that are able to contribute to the solutions of the managerial challenges in health care (Elkhuizen, van Sambeek, Hans, Krabbendam, & Bakkar, 2007; Koizumi, Kuno, & Smith, 2005).

When the U.S. Academy of Sciences reported from a committee's work where health care professionals together with engineers have identified systems engineering tools and technologies that could help the health system to overcome the poor performance in health care, they came up with a list where both queuing methods and simulations figured on the list (Reidd et al., 2005).

The health care sector as a whole has been reluctant to embrace these tools, even though they have been shown to yield valuable returns to the small but growing number of health care organisations and clinicians that have applied them. Relatively few health care professionals or administrators are equipped to think analytically about health care delivery as a system. Widespread use of systems-engineering tool therefore will require efforts such as the following from the Academy report:

- Incentives for health care providers to use system tools to improve the quality of care and the efficiency of care delivery
- Dissemination efforts by organizations that have used or promoted system-engineering tools in the health care delivery
- Support in the form of textbooks, instructional software, and other tools to train individual providers in the use of system-engineering tools
- Support for research to advance the application and utility of systems engineering in health care delivery
- Health care providers should ensure that current and future health care professionals have a basic understanding of how systems-engineering tools work and their potential benefits
- Support programs to educate and train present and future leaders and scholars in health care, engineering, and management in health systems engineering and management.

CONCLUSION

Although the case concerns a Danish hospital, it should be stressed that the problem of balancing the department capacities in a hospital is general. The numerous reports from using Operations Management in health care (Visser & Beech, 2005) show that even though modern hospitals around the world are not identical, they have very much in common. All over there are surgical and medical departments with various specialities, outpatient clinics, operations theatres, recovery rooms, intensive care units, and clinical service departments such as x-ray and laboratory. And they all face the challenge of balancing the capacity across departments in order to make a maximum flow of patients without local departmental bottlenecks. Therefore the use of SD illustrated by the Danish hospital case is relevant and applicable in hospitals in general.

An analysis, like the one described in this article, does require a lot of information and hard data about the daily operations. Clinicians often find it time consuming, and sometimes they doubt that it is necessary to go through all the data. Approaching the business of hospital employees from slightly new angles induces a natural scepticism among clinicians in particular. Especially when the data bring forward a picture of the daily operation that in some cases differs from their own mental models. This raises a lot of questions and time is needed to go through it thoroughly. If the analysis is aimed at achieving a common view of what can be improved and how, then those who will be responsible for implementing the recommendations from the analysis have to make sure that the analysis is based on the best information available. Therefore, very often, it does take time to bring out relevant data for this type of analysis.

The lack of available specific operational data also points out that something is missing in the management information and balanced scorecard at hospitals today. It is a little worrying to realise that many decisions about reorganising hospitals are made without the necessary operational data available. This local observation is a problem of a much larger proportion according to, for example, the American Academy of Science (Reid et al., 2005) that stress the lack of a system view in health care.

Regarding the use of models in balancing the surgical capacity in a hospital and many other important decisions to be made in health care, the clinicians' scepticism may be justified. Can these kinds of models be trusted? Should not the model be validated even further to make sure that it can handle all types of changes in a hospital? Our answer to a great extent follows Sterman (Sterman, 2000, p. 890),

All models are wrong, so no models are valid or verifiable in the sense of establishing the truth. The question facing clients and modellers is never whether a model is true but whether it is useful. The choice is never whether to use a model. The only choice is which model to use. Selecting the most appropriate model is always a value judgment to be made by reference to the purpose. Without a clear understanding of the purpose for which the model is to be used, it is impossible to determine whether you should use it as a basis for action.

System Dynamics is a general method for analysing and obtaining insight of systems and systems behaviour (Sterman, 2000). Therefore the tools (mapping and modelling), as mentioned, are very much applicable in other cases, where the flow of patients is in focus. The U.S. Academy of Science stresses that the health care sector is in need of system engineering tools in an effort to make improvement and System Dynamics is one the of tools recommended (Reidd et al., 2005).

What we seek when doing the modelling is consistency and the opportunity to evaluate the different proposals for improvement in broader terms rather than to derive an estimate of the expected effect with a probability of 95%.

As mentioned, it seems that the different concepts for organisational improvements have much in common when it comes to the tools and techniques for mapping. System Dynamics modelling techniques are presented as tools for making Business Process Reengineering (Spurr, Layzell, Jennison, & Richards, 1993), and there are similarities in the way we have used the ideas from System Dynamics and The Learning Organisation and the presentation of Lean Production (Bicheno, 2004).

Mapping is emphasized as an important tool in Lean. Nevertheless, Womack et al. (1990) (Rother & Shook, 2003) report that many well-intended initiatives get a poor result because the mapping exercise is skipped. A glance on some of the Lean references in the Health Care Sector (Jimmerson, Weber, & Sobek, 2004; Laursen, Gertsen, & Johansen, 2003; Lean), in our view, confirm the evaluation of Womack et al. (1990). There is room for improvements in the mapping part of Lean implementation. System Dynamics offers tools for mapping that integrate the visual clarification and mutual understanding of the system (value chain or patient flow) with the succeeding collection of data and modelling. This is an integration that secures a consistent use of organisational knowledge in suggesting and implementing proposals for improvements.

REFERENCES

Beirema, L. (2003). System thinking: A new lens for old problems. *The Journal of Continuing Education in the Health Professions, 23,* 27-33.

Bicheno, J. (2004). *New Lean toolbox* (Danish version). Id-tripe.

Bolling, W.D. (1972). Queuing model for a hospital emergency room. *Industrial Engineering,* 26-33.

deLuc, K. (2001). *Developing care pathways.* Radcliffe Medical Press.

Deming, W.E. (1982). *Out of the crisis.* MIT-CAES.

Elkhuizen, S.G., van Sambeek, J.R.C., Hans, E.W., Krabbendam, K., & Bakker, P.J.M. (2007). Applying the variety reduction principle to management of ancillary services. *Health Management Review, 32,* 37-45.

Gonzales-Busto, B., & Garcia, R. (1999). Waiting lists in Spanish public hospitals: A system dynamics approach. *System Dynamics Review, 15*(3), 201-224.

Hammer, M., & Champy, J. (1993). *Reengineering the coorporation: A manifesto for business revolution.* Harper.

Homer, J., Hirsch, G., Minniti, M., & Pierson, M. (2004). Models for collaboration: How system dynamics helped a community organize cost-effective care for chronic illness. *System Dynamics Review, 20*(3), 199-222.

Hsieh, P.-J., Tan, B., & Lin, C. (2003). An integrated approach to determine hospital outpatient staffing needs. *International Journal of Health care Technology and Management, 5*(1/2), 96-122.

Jimmerson, C., Weber, D., & Sobek, D.K. (2004). *Reducing waste and errors: Piloting lean principles at IHC.* Retrieved February 6, 2008, from www.lean.org

Johanneson, H.J., McHugh, P., Pendlebury, A.J., & Wheeler, W.A., III (1993). *Business process reengineering: Breakpoint strategies for market dominance.* Wiley.

Koizumi, N., Kuno, E., & Smith, T.S. (2005). Modelling patient flows using a queuing network with blocking. *Health Care Management Science, 8,* 49-60.

Lane, D. (2000). Diagramming conventions in system dynamics. *Journal of the Operational Research Society, 51,* 241-245.

Lane, D.C., Monefeldt, C., & Lund, C.M. (2000). Looking at the wrong place for health care improvements: A system dynamics study of an accident and emergency department. *Journal of the Operational Research Society, 51,* 518-531.

Laursen, M.L., Gertsen, F., & Johansen, J. (2003). *Applying lean thinking in hospitals - exploring implementing difficulties.* Center for Industrial Production, Aalborg University. Denmark.

Lean. Retrieved February 6, 2008, from www.lean.org.

Nolan, T. (1998). Understanding medical systems. *Annals of Internal Medicine, 128*(4), 293-298.

Reidd, P., Compton, D., Grossmann, J., & Fanjiang, G. (2005). *Building a better delivery system: A new engineering/health care partnership.* National Academies Press.

Rother, M., & Shook, J. (2003). *Learn to see.* The Lean Enterprise Institute. Brookline.

Schmitz, H.H., & Kwak, N.K. (1972). Monte Carlo simulation of operating room usage. *Operations Research,* 1171-1180.

Senge, P. (1990). *The fifth dicipline. The art and pracice of the learning organisation.* New York: Doubleday.

Spurr, K., Layzell, P., Jennison, L., & Richards, N. (1993). *Software assistence for business reengineering.* Wiley.

Sterman, J. (2000). *Business dynamics: System thinking and modelling for a complex world.* McGraw-Hill.

Van Ackere, A., & Smith, P. (1999). Towards a macro model of the national health service waiting list. *System Dynamics Review, 15*(3), 225-252.

Visser, J., & Beech, R. (2005). *Health operations management - patient flow logistics in health care.* London: Routledge.

Waaler, H., & Iversen, T. (1998). Køer som resultat av statistiske tilfeldigheder. *Tidskrift for Norsk Lægeforening (Norwegian), 118*(30), 4742-4745.

Wolstenholme, E. (1999). A patient flow perspective of U.K. Health Services: Exploring the case for new "intermediate care" initiatives. *System Dynamics Review, 15*(3), 253-271.

Womack, J.P., Jones, D.T., & Roos, D. (1990). *The machine that changed the world*. Macmillan/Rawson Associates.

KEY TERMS

Balancing Health Care Capacity: Coordinating the capacities for the operational units involved in flow of patient in order to avoid bottlenecks and to secure a high rate in capacity usage.

Simulation: Calculating systems or health care organizations operational behaviour based on a mathematical model describing the system and with varying values for central parameters in the model.

System Dynamics: A methodology for studying and managing complex feedback systems, such as one finds in business and other social systems.

System View: An expression that is meant to remind you that you should focus on the whole system and to prevent you from (sub-) optimizing parts of it.

Visualization: Making maps showing the patient flows and central variables controlling the flows. Through the maps, different actors in the health care organizations elicit their common mental model of the system under consideration.

B

Biomedical Data Warehouses

Jérôme Darmont
University of Lyon 2, France

Emerson Olivier
University of Lyon 2, France

INTRODUCTION

With the growing use of new technologies, health care nowadays is undergoing significant changes. The development of electronic health records will indeed help enforce personalized, lifetime health care and presymptomatic treatment, as well as various analyses over populations of patients. Information-based medicine has to exploit medical decision-support systems and requires the analysis of various, heterogeneous data, such as patient records, medical images, biological analysis results, and so forth (Saad, 2004).

Data warehousing technologies (Inmon, 2002; Kimball & Ross, 2002) are now considered mature and can form the base of such a decision-support system. Although data warehousing primarily allows the analysis of numerical data, its underlying concepts remain valid for what we term *complex data*. To summarize, data may be qualified as complex if they are (Darmont, Boussaïd, Ralaivao & Aouiche, 2005):

- **Multiformat:** Represented in various formats (databases, texts, images, sounds, videos) and/or
- **Multistructure:** Diversely structured (relational databases, XML document repositories, file collection) and/or
- **Multisource:** Originating from several different sources (distributed databases, the Web) and/or
- **Multimodal:** Described through several channels or points of view (radiographies and audio diagnosis of a physician, data expressed in different scales or languages) and/or
- **Multiversion:** Changing in terms of definition or value (temporal databases, periodical surveys with evolving items)

In this context, the warehouse measures (although not necessarily numerical) remain the indicators for analysis, and analysis is still performed following different perspectives represented by dimensions. Large data volumes and their dating are other arguments in favor of this approach (Darmont, Boussaïd, Bentayeb, Rabaseda & Zellouf, 2003). Data warehousing also can support various types of analysis, such as statistical reporting, online analysis (OLAP), and data mining.

The aim of this chapter is to present an overview of the existing data warehouses for biomedical data and to discuss the issues and future trends in biomedical data warehousing. We illustrate this topic by presenting the design of an innovative, complex data warehouse for personal, anticipative medicine.

BACKGROUND

The first family of medical data warehouses we identify are repositories tailored for supporting data mining (Miquel & Tchounikine, 2002; Prather, Lobach, Goodwin, Hales, Hage & Hammond, 1997; Sun, Huang, Horng, Huang & Tsou, 2004; Tchounikine, Miquel & Flory, 2001). However, since data mining techniques take attribute-value tables as input, some of these warehouses (Prather et al., 1997; Sun et al., 2004) do not bear the multidimensional, starlike architecture that is typical in data warehouses. This modeling choice precludes OLAP navigation and is not very evolutionary; new analysis axes, or dimensions, cannot be easily plugged into the warehouse.

The most "canonical" medical data warehouse among these proposals is a cardiology data warehouse (Miquel & Tchounikine, 2002; Tchounikine et al., 2001). Its aim is to ease medical data mining by integrating data and processes into a single warehouse. However, raw sensor data (e.g., electrocardiograms) are stored separately from multidimensional data (e.g., patient identity, therapeutic data), while it might be interesting to integrate them all.

A second family is constituted of biological data warehouses that focus on molecular biology and genetics (Engström & Asthorsso, 2003; Eriksson & Tsuritani, 2003; Schönbach, Kowalski-Saunders & Brusic, 2000; Shah, Huang, Xu, Yuen, Ling & Ouellette, 2005; Sun et al., 2004), and bear interesting characteristics. For instance, some of them include metadata and ontologies from various public sources such as RefSeq or Medline (Engström & Asthorsso, 2003; Shah et al., 2005). The incremental maintenance and evolution of the warehouse is also addressed (Engström & Asthorsso, 2003). However, the particular focus of these approaches makes them inappropriate to more general needs, which may be both different and much more diversified.

Eventually, Boussaïd, Ben Messaoud, Choquet, and Anthoard (2006) recently proposed an eXtended Markup Language (XML) based methodology named X-Warehousing for warehousing complex data. This approach has been applied on a corpus of patient records extracted from the Digital Database for Screening Mammography[1]. The warehouse is a collection of XML documents representing OLAP facts that describe suspect areas in mammographies. It aims at breast cancer computer-aided diagnosis.

A COMPLEX DATA WAREHOUSE FOR PERSONALIZED, ANTICIPATIVE MEDICINE

Context and Motivation

Dr. Jean-Marcel Ferret, former physician of the French national soccer team, is the promoter of the personalized and anticipative medicine project. His aim is to extend results and empirical advances achieved for high-level athletes (not only soccer players) to other populations and to make the analyzed subjects the managers of their own health capital by issuing recommendations regarding, for example, life style, nutrition, or physical activity. This is meant to support personalized, anticipative medicine. In order to achieve personalized, lifetime health care and presymptomatic treatment, a decision-support system must allow transverse analyses of given populations of patients and the storage of global medical data such as biometrical, biological, cardiovascular, clinical, and psychological data. It also must be evolutionary in order to take into account future advances in medical research. More precisely, such a system must be able to store complex medical data and allow quite different kinds of analyses to support the following:

1. Personalized and anticipative medicine (in opposition to curative medicine) for well-identified patients;
2. Broadband statistical studies over given populations of patients.

We selected a data warehousing approach to answer this need. A data warehouse can indeed support statistical reporting and cross-analyses along several dimensions. Furthermore, dated personal data can be stored to propose full patient records to physician users. However, data complexity must be handled. For instance, multimedia documents such as echocardiograms must be stored and explicitly related to more classical medical data such as the corresponding patient or diagnoses. Users must be able to display and exploit such relationships, either manually (which is currently the case) or automatically (here, we anticipate the advances of multimedia mining and the development of advanced OLAP operators). The existing research proposals from the literature do not fulfill this requirement. In particular, Tchounikine, et al. (2001) store raw sensor data separately from multidimensional data, only as an archive. Finally, Dr. Ferret had quite an immediate need for an operational, efficient system. Hence, we chose to rely on the efficacy of a relational implementation. Although XML warehousing is particularly appropriate to complex, medical data, the performance of native XML Database Management Systems (DBMSs) is indeed currently not satisfactory for warehousing purposes, both in terms of storage capacity and response time.

In the remainder of this section, we present the global architecture of our medical data warehouse, two examples of simple datamarts (i.e., datamarts that store "simple" data), an example of complex datamart (i.e., storing complex data), as well as implementation issues.

Global Architecture

Our data warehouse is organized as a collection of interconnected datamarts sharing common dimensions. Each datamart stores the data related to a given medical field (e.g., biological analysis results, biometrical data, cardiovascular data, psychological data). These data are described by dimensions. Some are common to all the datamarts (e.g., patient information, dates), and some are specific to a given datamart. The warehoused data originate from diverse providers and are very heterogeneous (e.g., qualitative or numerical data, texts, medical images).

To make our solution evolutionary, we adopted a bus architecture (Firestone, 2002; Kimball & Ross, 2002) for our data warehouse. It is composed of a set of conformed dimensions and standardized definitions of facts. In this framework, the warehoused data related to every medical field we need to take into account represent datamarts that are plugged into the data warehouse bus and receive the dimension and fact tables they need. The union of these datamarts may be viewed as the whole data warehouse (Kimball & Ross, 2002).

Using conformed dimensions (i.e., dimensions that have the same meaning and the same modeling in all the fact tables to which they are linked) helps to guarantee that datamarts are not "stovepipes" that are independent from one another. Hence, when a new aspect of health care (e.g., chiropody) needs to be included in the warehouse, it is easier to plug it into the data warehouse bus.

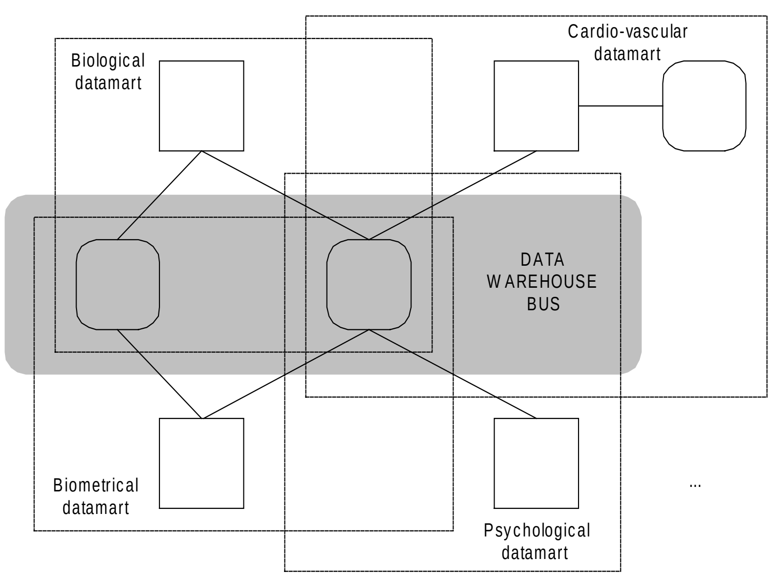
Figure 1 represents the global architecture of our data warehouse. Straight squares symbolize fact tables, round squares symbolize dimensions, dotted lines embed the different datamarts, and the data warehouse bus bears a gray background. It is constituted by dimensions that are common to several datamarts (and thus, fact tables). The main dimensions that are common to all our datamarts are patient, data provider, time, and medical analysis (that regroups several kinds of analyses) (see the next section). Of course, some datamarts (e.g., the cardiovascular datamart) do have specific dimensions that are not shared.

Simple Datamarts: The Biological and Biometrical Datamarts

Input biological data are actually the results of various biological examinations (e.g., biochemistry, protein checkup, hematology, immunology, toxicology, etc.), which are themselves subdivided into actual analyses (e.g., a hematology examination consists in reticulocyte numbering and a hemogram). These data are available under the form of unnormalized spreadsheet files from various sources. They are thus heterogeneous and often refer to the same examination using different terms or abbreviations, use different units for the same numerical values, and so forth. This heterogeneity is dealt with during the ETL (Extract, Transform, and Load) process (see the following implementation section).

Biometrical data are measured during medical examinations. They include weight, height, pulse rate, fat percentage, and blood pressure. Although their structure

Figure 1. Data warehouse global architecture



is simpler than that of biological data, they require a fine granularity that must be taken into account. For example, the weight of an athlete may be measured twice a day, before and after practice. This has an impact on data warehouse modeling. More precisely, it helps in defining the granularity of the time dimension hierarchy (see the following).

Figure 2 represents the architecture of our biological and biometrical datamarts. The biological fact table stores an exam result under the form of a numerical value (e.g., reticulocyte numbering). It is described by four dimensions: patient, time of the examination, data provider (typically the laboratory performing the analysis), and the analysis itself. The biometrical fact table stores numerical biometrical values (e.g., weight). It is described by the same dimensions as the biological datamart, the "analysis" actually representing a measurement. The patient, data provider, time, and medical examination dimensions are thus all shared, conformed dimensions. Attributes in dimension tables are not detailed due to confidentiality constraints.

Note that the medical analysis and time dimensions are further organized into hierarchies to take into account the particularities identified into the source data. Here, biological and biometrical data are distinguished; the simple biometrical data are not normalized nor organized in a hierarchy, while the biological data are. Hence, the description of biometrical facts only appears in the medical analysis dimension table, while biological facts are described further by a hierarchy of biological examinations and categories.

Each datamart is thus modeled as a snowflake schema rather than a simpler, classical star schema. Since the biological and biometrical fact tables share their dimensions, our overall data warehouse follows a constellation schema. In such an architecture, it is easy to add in new fact tables described by existing dimensions.

Finally, although this is not indicated in Figure 2, our data warehouse also includes metadata that help in managing both the integration of source data into the warehouse (e.g., correspondence among various labels or numerical units, the French SLBC biomedical nomenclature, etc.) and their exploitation during analysis (e.g., the interval between which an examination result is considered normal).

Complex Datamart: The Cardiovascular Datamart

Figure 3 represents the architecture of our cardiovascular datamart. The complex nature of source data, which are constituted of raw measurements (e.g., ventricle size), multimedia documents (e.g., echocardiograms), and a conclusion by a physician, cannot be embedded in a single standard fact table. Hence, we actually exploit a set of interrelated tables that together represents the facts, represented as dotted, straight squares in Figure 3.

Figure 2. Biological and biometrical datamarts' architecture

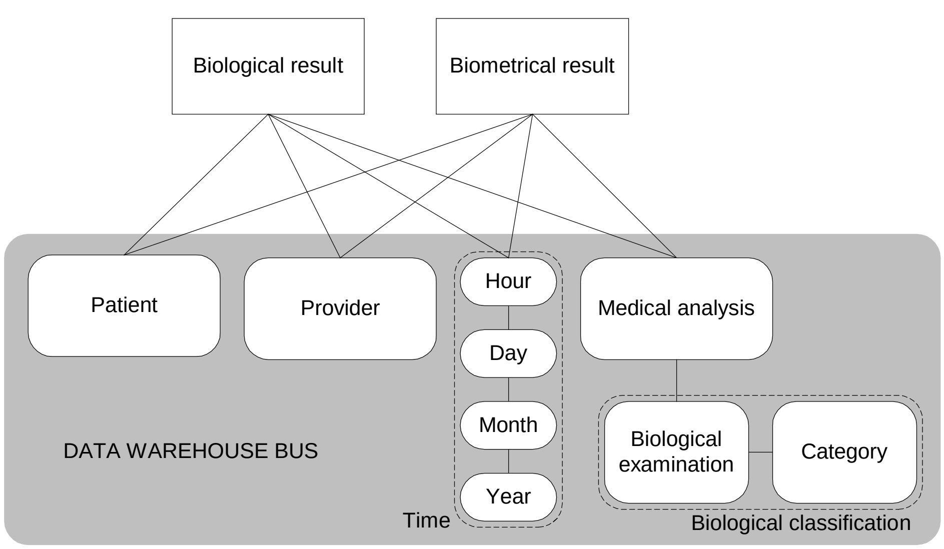
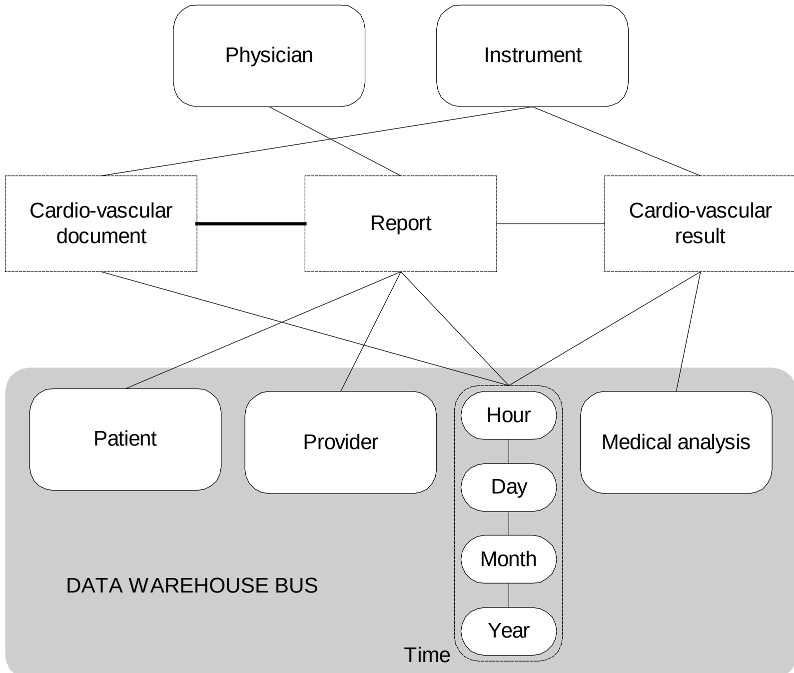


Figure 3. Cardiovascular datamart's architecture



The report mainly contains the physician's conclusion. It is the central element in our "complex fact." It is linked to several numerical analysis results that help build the physician's conclusion. It is also related to multimedia documents such as medical images, which also help in devising the diagnosis. Note that this many-to-many relationship is represented as a bold line in Figure 3. Some documents may indeed be referred to by several reports; for instance, to take into account a patient's evolution over time, through distinct series of echocardiograms. Each component of our "complex fact" may be linked individually to the dimensions. Cardiovascular documents and results are indeed not descriptors of the would-be fact; they are part of a fuzzier fact that is composed of several entities.

Finally, note that cardiovascular documents currently cannot be exploited by OLAP technologies. However, we need to store them and have them accessible for medico-legal reasons. Furthermore, extensions to the usual OLAP operators should make this exploitation possible in the near future.

Implementation

Our data warehouse is meant to be used on an intranet. Hence, we developed a prototype using popular tools that are freely available; namely, the Apache Web server that implements the Secure Sockets Layer (SSL) protocol, which is important for the security of our solution; the MySQL DBMS that is fast and reasonably reliable; and the powerful PHP Web scripting language. This software configuration is sufficient for our prototype and very light and easy to deploy, but we can easily switch one component for a more robust one, if necessary. This is especially true for the DBMS component, since our data warehouse will have to store voluminous complex data.

At the ETL level, in order to achieve the integration of biological data into our data warehouse (biometrical data are much simpler to integrate), we developed a series of scripts that help in extracting the data from the source spreadsheet files, in transforming them so they are consistent with the data already present in the warehouse, and eventually in loading them into the data warehouse. This is currently only a semi-automatic process that proceeds in three steps/PHP scripts:

1. Remove doubles in the input files;
2. Check whether dimension data exist in the data warehouse and create them if not;
3. Transform the data whenever necessary (measure units, dimension key correspondence), and finally, load the data into the warehouse.

Eventually, the first analyses we implemented over our data warehouse are divided into two kinds that correspond to the goals we mentioned earlier: patient-oriented results and statistical analyses over groups of patients. For the first kind of output, our front-end tool includes a patient search engine; the display of a given patient's biological and biometrical data following the biological or biometrical examination dimension, respectively; the visual identification of out-of-norm values; and graphical representations of the evolution of any numerical measure over time.

To achieve the second kind of output, we implemented the notion of group of patients. The user can create, modify, and suppress such groups (e.g., soccer players, back and forward players, etc.); select group members according to various criteria; and eventually compare the various groups through numerical tables and graphs.

FUTURE TRENDS

The need for medical decision-support systems has grown very strong nowadays, and many such systems have been developed. However, most of them store, process, and allow the analysis of "simple" data. Taking complex data into account is now a pressing challenge (Saad, 2004).

We propose in this chapter a solution for warehousing (i.e., storing complex data). Many research perspectives still lie ahead, though. The multimedia data we store are indeed available for display, but we cannot truly analyze them. Developing OLAP and/or data mining techniques that operate on complex data is currently becoming a research field in itself. We could consider, for instance, medical images as a dimension in a medical warehouse. Aggregating such data with clustering-based OLAP operators (Ben Messaoud, Rabaséda, Boussaïd & Bentayeb, 2004) could help in rolling up and drilling down along this dimension.

To achieve this goal, integrating semantic information about the processed complex data is mandatory, and OLAP and data mining techniques must rely heavily on metadata and domain-specific knowledge. The integration of this knowledge into complex data warehouses and, most of all, its exploitation for data analysis, are also exciting challenges.

CONCLUSION

We have presented in this chapter an overview of biomedical data warehousing, particularly focused on the complex data warehouse for personal, anticipative medicine we propose. The aim of this tool is to support both patient-oriented analyses and broadband statistical studies over populations of patients. By adopting a data warehouse bus architecture, we designed our system to be global to take into account several medical fields, and evolutionary to take into account future advances in medical research. We also have briefly presented the prototype we implemented to achieve our analysis goals.

The direct perspectives over this work are twofold. The first one concerns the actual contents and significance of the data warehouse. It involves modeling and adding in new datamarts. The psychology datamart from Figure 1 is already implemented, and more are either currently developed (such as a medical background datamart), being developed, or scheduled. This helps in broadening the scopes of analyses. Other output than statistical reports also are envisaged. Since we adopted a dimensional modeling, OLAP navigation is also definitely possible, and "attribute-value" views easily could be extracted from the data warehouse to allow data mining explorations.

The second kind of perspective is more technical and aims at improving our prototype. This includes automating and generalizing the ETL process on all the datamarts, which is currently an ongoing task. We also follow other leads to improve the user-friendliness of our interfaces and the security of the whole system, which is particularly primordial when dealing with personal medical data.

REFERENCES

Ben Messaoud, R., Rabaséda, S., Boussaïd, O., & Bentayeb, F. (2004). *OpAC: A new OLAP operator based on a data mining method*. Proceedings of the *Sixth International Baltic Conference on Databases and Information Systems (DB&IS 04), Riga, Latvia*.

Boussaïd, O., Ben Messaoud, R., Choquet, R., & Anthoard, S. (2006). *X-warehousing: An XML-based approach for warehousing complex data*. Proceedings of the *10th East-European Conference on Advances* *in Databases and Information Systems (ADBIS 06), Thessaloniki, Greece. LNCS.* 4152, 39-54.

Darmont, J., Boussaïd, O., Bentayeb, F., Rabaseda, S., & Zellouf, Y. (2003). Web multiform data structuring for warehousing. *Multimedia Systems and Applications, 22,* 179-194.

Darmont, J., Boussaïd, O., Ralaivao, J.C., & Aouiche, K. (2005). *An architecture framework for complex data warehouses*. Proceedings of the *7th International Conference on Enterprise Information Systems (ICEIS 05), Miami, Florida,* 370–373.

Engström, H., & Asthorsso, K. (2003). *A data warehouse approach to maintenance of integrated biological data*. Proceedings of the Workshop on Bioinformatics, 19th International Conference on Data Engineering (ICDE 03), Bangalore, India.

Eriksson, T., & Tsuritani, K. (2003). *A dimensional data warehouse for biological data*. Proceedings of the 11th International Conference on Intelligent Systems for Molecular Biology (ISMB 03), Brisbane, Australia.

Firestone, J.M. (2002). DKMA and the data warehouse bus architecture. DKMS Brief. 7. Retrieved from http://www.dkms.com/papers/dkmadwbus.pdf

Inmon, W.H. (2002). *Building the data warehouse*. John Wiley & Sons.

Kimball, R., & Ross, M. (2002). The data warehouse toolkit. John Wiley & Sons.

Miquel, M., & Tchounikine, A. (2002). *Software components integration in medical data warehouses: A proposal*. Proceedings of the 15th IEEE Symposium on Computer-Based Medical Systems (CBMS 02), Maribor, Slovenia, 361–364.

Prather, C., Lobach, D.F., Goodwin, L.K., Hales, J.W., Hage, M.L., & Hammond, W.E. (1997). *Medical data mining: Knowledge discovery in a clinical data warehouse*. Proceedings of the AMIA Annual Fall Symposium, Nashville, Tennessee.

Saad, K. (2004). *Information-based medicine: A new era in patient care*. Proceedings of the ACM 7th International Workshop on Data Warehousing and OLAP (DOLAP 04), Washington, D.C, 58.

Schönbach, C., Kowalski-Saunders, P., & Brusic, V. (2000). Data warehousing in molecular biology. *Briefings in Bioinformatics, 1*(2), 190–198.

Shah, S.P., Huang, Y., Xu, T., Yuen, M.M., Ling, J., & Ouellette, B.F. (2005). Atlas—A data warehouse for integrative bioinformatics. *BMC Bioinformatics, 6*(1), 34.

Sun, Y.M., Huang, H.D., Horng, J.T., Huang, S.L., & Tsou, A.P. (2004). *RgS-Miner: A biological data warehousing, analyzing and mining system for identifying transcriptional regulatory sites in human genome*. Proceedings of the 15th International Database and Expert Systems Applications Conference (DEXA 04), Zaragoza, Spain. LNCS. 3180, 751–760.

Tchounikine, A., Miquel, M., & Flory, A. (2001). *Information warehouse for medical research*. Proceedings of the 3rd International Conference on Data Warehousing and Knowledge Discovery (DaWaK 01), München, Germany. LNCS. 2114, 208–218.

KEY TERMS

Bus Architecture: Set of conformed dimensions and standardized definitions of facts (Kimball & Ross, 2002). Datamarts "plug into" this bus to receive the dimensions and facts they need (Firestone, 2002).

Complex Data: Data that are not numerical or symbolic (e.g., multimedia, heterogeneous data stored on multiple platforms).

Data Warehouse: Subject-oriented, integrated, time-variant and nonvolatile collection of data in support of management's decision-making process (Inmon, 2002).

Datamart: Logical and physical subset of the overall data warehouse, usually dedicated to a given activity.

ETL: Data warehousing process that includes extracting data from external sources, transforming them and finally loading them into the warehouse.

Information-Based Medicine: Utilizing information technology to achieve personalized health care (Saad, 2004).

OLAP (Online Analytical Processing): An approach for processing decision-support, analytical queries that are dimensional in nature.

ENDNOTE

[1] Refers to http://marathon.csee.usf.edu/mammography/database.html

Biomedical Signal Compression

Pedro de A. Berger
University of Brasilia, Brazil

Francisco A. de O. Nascimento
University of Brasilia, Brazil

Leonardo R. A. X. de Menezes
University of Brasilia, Brazil

Adson F. da Rocha
University of Brasilia, Brazil

Joao L. A. Carvalho
University of Southern California, USA

INTRODUCTION

Digitization of biomedical signals has been used in several areas. Some of these include ambulatory monitoring, phone line transmission, database storage, and several other applications in health and biomedical engineering. These applications have helped in diagnostics, patient care, and remote treatment. One example is the digital transmission of ECG signals, from the patient's house or ambulance to the hospital. This has been proven useful in cardiac diagnoses.

Biomedical signals need to be digitally stored or transmitted with a large number of samples per second, and with a great number of bits per sample, in order to assure the required fidelity of the waveform for visual inspection. Therefore, the use of signal compression techniques is fundamental for cost reduction and technical feasibility of storage and transmission of biomedical signals.

The purpose of any signal compression technique is the reduction of the amount of bits used to represent a signal. This must be accomplished while preserving the morphological characteristics of the waveform. In theory, signal compression is the process where the redundant information contained in the signal is detected and eliminated. Shannon (1948) defined redundancy as "the fraction of unnecessary information, and therefore repetitive in the sense that if it was missing, then the information would still be essentially complete, or it could at least be recovered."

Signal compression has been widely studied during the past decades, and several references discuss this subject (Gersho & Gray, 1992; Jayant & Noll, 1982; Sayood, 1996).

Signal compression techniques are commonly classified in two categories: lossless and lossy compression. Lossless compression means that the decoded signal is identical to the original one. In lossy compression, a controlled amount of distortion is allowed. Lossy signal compression techniques show higher compression gains than lossless ones.

BACKGROUND

Lossless Compression

Lossless signal compression techniques are less efficient with respect to compression gains. They may be used in combination with lossy compression techniques, especially in cases where the maximum allowed distortion has been reached, and additional compression is needed. Among several lossless compression techniques, we highlight entropy coding (Gersho & Gray, 1992), Run-Length, Huffman (Huffman, 1952), arithmetic coding (Witten, 1987), and delta coding (Ken, 1985).

Run-Length Coding

Data files frequently present sequentially repeated characters (a character run). For instance: text files use several spaces to separate sentences and paragraphs. Digital signals may contain the same value, or the same character representing that value in its data file, repeated many times sequentially. This indicates that the signal is not changing, as in the isoelectric segments of ECG signals, for example.

Figure 1 shows an example of Run-Length coding of a data set that contains runs of zeros. Each time the coder finds a zero in the entry data, two values are written in the output data. The first of these values is a zero indicating that the Run-Length codification started. The second value is the amount of zeros in the sequence. If the run of zeros in the input data set is in average larger than two, then the Run-Length coder will achieve data compression.

Huffman Coding

In Huffman coding, the data are represented as a set of variable length binary words. The lengths depend on the occurrence frequency of the symbols used for representing each signal value. Characters that are used often are represented with fewer bits, and those that are seldom used are represented with more bits.

Figure 2 shows an example of how a Huffman code is generated, given a data set X and its characteristic probability of occurrence—p(X). The character codes are generated by combining the bits of a tree with ramifications, adding their probabilities and restarting the process until only one character remains. This process generates a three with ramifications linked to bits 0 and 1. The codes for each character are taken in the inverse path of these ramifications. Note that initial character arrangement is not relevant. In this example, we chose to encode the upper ramifications with bit 0 and the lower ones with bit 1. However, the opposite representation could have been used as well. Any decision criteria may be used in ramifications with equal probabilities.

Huffman coding has the disadvantage of assigning an integer number of bits to each symbol. This is a suboptimal strategy, because the optimal number of bits per symbol depends on the information content, and is generally a rational number. Arithmetic coding is a more sophisticated compression technique, based on Huffman coding concepts. It results in compression gains closer to the theoretical limits. In this coding, character sequences are represented by individual codes depending on their probability of occurrence. Huffman and Arithmetic codes are often used in combination with Run-Length coding, for further compressing biomedical signals that were first compressed using orthogonal transforms.

Delta Coding

Delta coding refers to signal compression techniques that store a digital signal as the difference between successive samples. Figure 3 shows an example of how this is performed. The first sample in the encoded signal is equal to the first sample of the original signal. Subsequent encoded samples are equal to the difference between the current sample and the previous sample of the original signal.

Using this technique, the encoded signal has a smaller amplitude dynamic range than the original signal. Therefore, it takes fewer bits to store or transmit the encoded signal. Delta coding is a particular case of Differential Pulse Code Modulation (DPCM). DPCM is used in combination with Huffman coders in several biomedical signal compression algorithms.

Figure 1. Run-length coding example

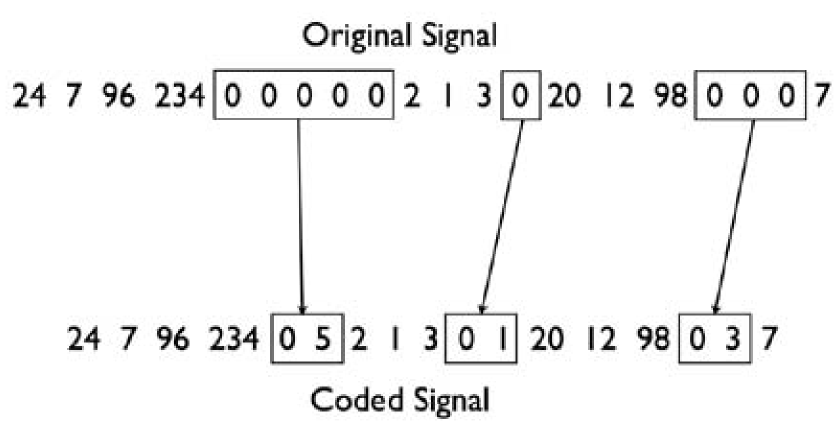
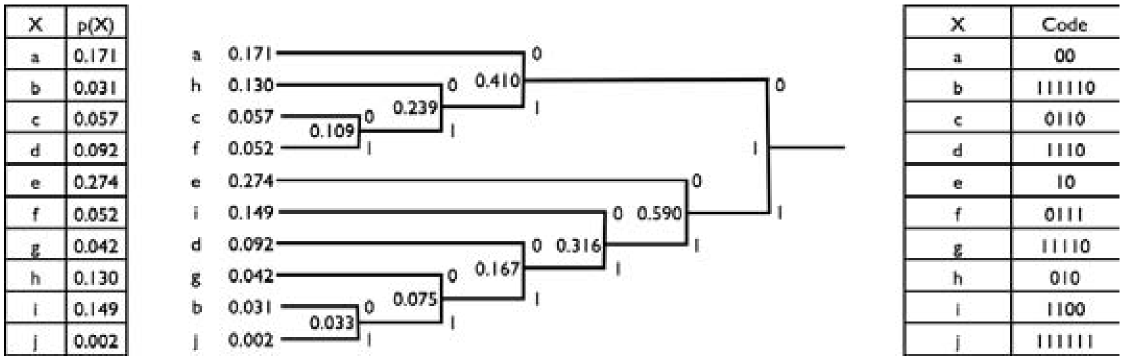


Figure 2. Huffman code construction example



X	p(X)
a	0.171
b	0.031
c	0.057
d	0.092
e	0.274
f	0.052
g	0.042
h	0.130
i	0.149
j	0.002

X	Code
a	00
b	111110
c	0110
d	1110
e	10
f	0111
g	11110
h	010
i	1100
j	111111

Figure 3. Delta coding example

Original Signal: 23 20 19 21 11 75 79 84 ...

Coded Signal: 23 -3 -1 2 -10 64 4 5 ...

Lossy Compression

There are two major categories of lossy compression techniques used with biomedical signals: direct methods, and transform compression.

Direct Methods

Direct methods encode signals in time domain. These algorithms depend on the morphology of the input signal. In most cases, these methods are complex, and provide lower compression efficiency than transform coders (discussed in the next section).

Direct compression methods use sophisticated processes and "intelligent" signal decimation and interpolation. In other words, these methods extract K "significant" samples of the original signal x(n), such that:

$$(n, x(n)),\ n = 0,\ldots, N-1 \rightarrow (n_k, x(n_k)),\ k = 0,\ldots,K-1, \qquad (1)$$

where k<N. The reconstruction of values between significant samples is performed by interpolation, using the generic expression (Sörnmo & Laguna, 2006):

$$\hat{x}(n) = \begin{cases} x(n), & n = n_0,\ldots,n_{k-1}; \\ f_{n_0,n_1}(n), & n = n_0+1,\ldots,n_1-1; \\ \vdots & \vdots \\ f_{n_{K-2},n_{K-1}}(n) & n = n_{K-2}+1,\ldots,n_{K-1}-1; \end{cases} \qquad (2)$$

The process of selecting significant samples is based on the characteristics of the signal, and on a tolerance interval criterion for the reconstruction error. The interpolation function $f_{n_{k-1},n_k}(n)$ is generally a low order polynomial, of either zero or first order. These polynomials approximate the original signal with a series of lines connecting the K significant samples. The geometric parameters of these lines (length and inclination) are stored instead of the discarded sample values, resulting in a reduction in the amount of stored or transmitted data.

Transform Compression

Among the several methods for signal compression, the techniques based on transforms yield the best performance in terms of compression gain and fidelity of the waveform of the reconstructed signal.

Given a vector of data $x=[x(1)\,x(2)....\,x(N)]$, we can define an orthogonal transform as a linear operation given by a linear transformation T, such that:

$$y = Tx, \tag{3}$$

where $y=[y(1)\ y(2)\ ...\ y(N)]$ represents a vector of transformed coefficients, and T satisfies the orthogonality condition:

$$T^t = T^{-1} \tag{4}$$

Transform compression is based on a simple premise: when the signal is processed by a transform, the signal energy (information) that was distributed among all the samples in the time domain can be represented with a small number of transformed coefficients.

This can be observed in Figure 4, where an ECG signal is shown along with the correspondent coefficients in transformed domain. In this example, we use the discrete cosine transform (DCT). The DCT is used in the most popular standard for image compression, the *Joint Picture Expert Group* (JPEG).

Note that, in the transformed domain, the energy of the signal is concentrated in a small number of coefficients with high amplitude. Thus, if we store the high-amplitude coefficients, and discard the low-amplitude coefficients, the signal may be represented accurately enough, and can be recovered by inverse transformation. Figure 5 shows the same ECG signal in Figure 4, and the reconstructed signal after discarding 80% of the DCT coefficients. Note that, by storing only 20% of the DCT coefficients it is possible to reconstruct the original signal without visible distortions.

Currently, the most widely used transform for encoding biomedical signals is the discrete wavelet transform (Daubechies, 1988; Mallat, 1989). The wavelet transform is used for image compression—such as in the JPEG 2000 compression standard—and in more recent studies on biomedical signal compression. In this transform, a signal containing N samples is filtered by a pair of filters that decompose the signal into low- (L) and a high- (H) frequency bands. Each band is undersampled by a factor of two; that is, each band contains N/2 samples. With the appropriate filter design, this action is reversible. This procedure can be extended for two-dimensional signals, such as images. In Figure 6, we show an example of wavelet decomposition for a gray-scale image with 256 x 256 pixels. Similarly to what was observed using the DCT, many of the coefficients in the high-frequency subbands have amplitudes close to zero (very dark pixels), and it is possible to compress the image by discarding them.

Wavelet transform compression is evolving with respect to the way in which the coefficients are encoded. When a coefficient in a low-frequency subband is nonzero, there is a high probability that, at positions that correspond to high frequencies, the coefficients are also nonzero. Thus, the nonzero coefficients can be represented in a tree, beginning with a low frequency root. Figure 7 illustrates this concept. A single coefficient in the LL band of layer 1 has a correspondent coefficient in the other bands. The positions of the coefficients in layer 1 are mapped into four daughter-positions in each subband of layer 2.

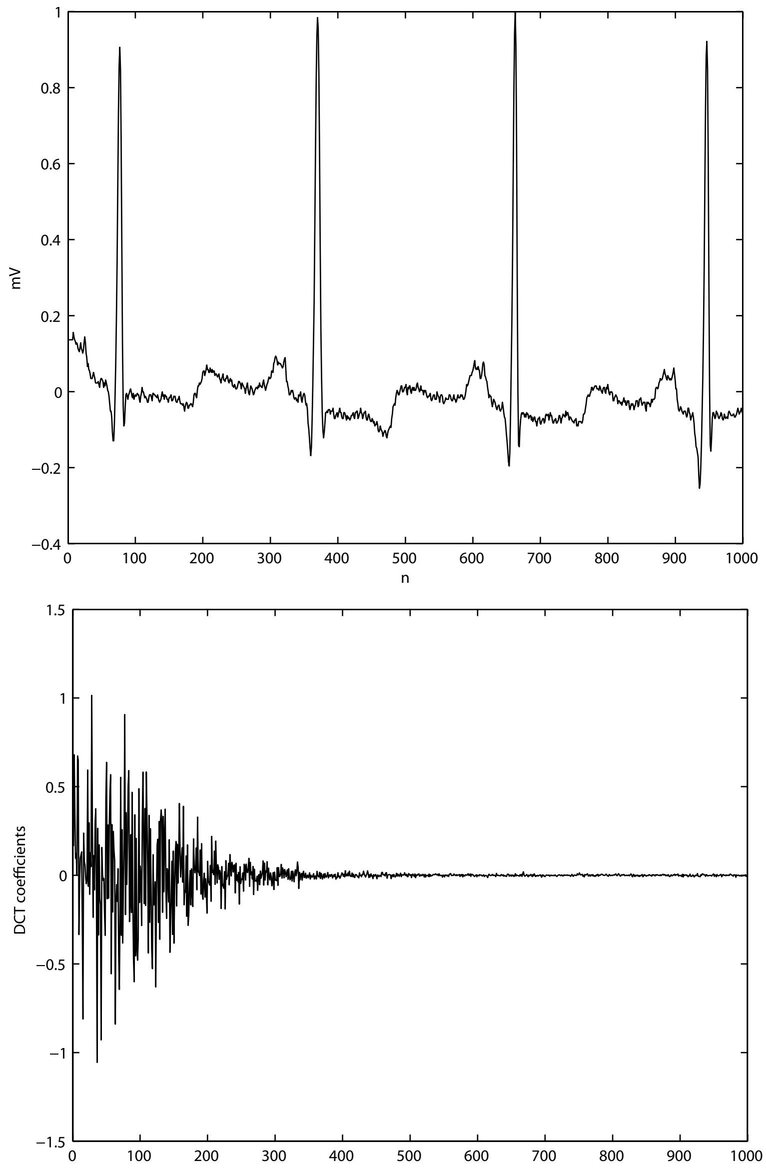
An efficient way of coding the coefficients that are non-zero is to code each tree of coefficients, beginning with the root decomposition level. The coefficients at the lower layers are encoded, and followed by their children coefficients in the higher layer, until a null coefficient is found. The next coefficients of the tree have great probability of also being null, and they are replaced by a code that identifies a tree of zeros (zerotree code). This method is called Embedded Zerotree Wavelet (EZW). Another method, similar to the EZW, and commonly used in the wavelet coefficients encoding, is the Set Partitioning In Hierarchical Tree (SPIHT) (Said & Pearlman, 1996).

COMMENTS ON CURRENT AND FUTURE RESEARCH

The compression of biomedical signals has been extensively studied by the scientific community. In this article, we discuss more thoroughly the compression of electrocardiographic (ECG) and electromyographic (EMG) signals.

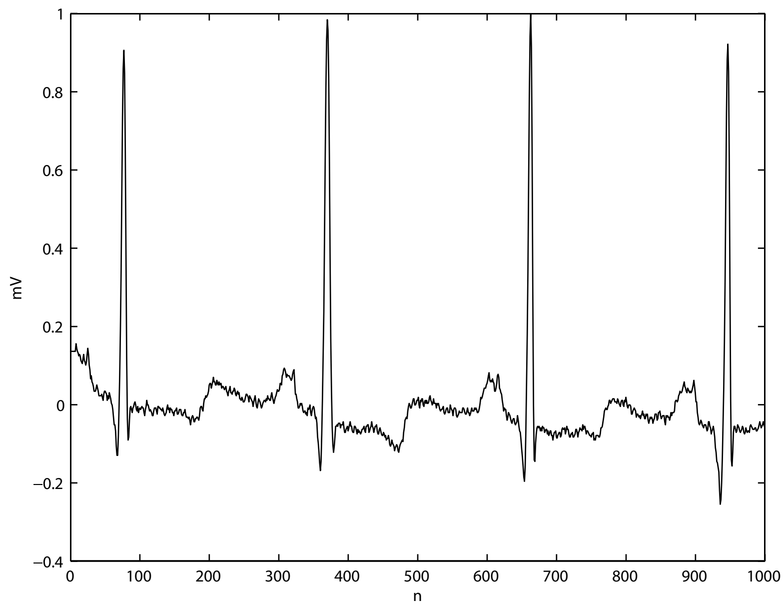
In general, we can group the techniques for compression of ECG signals in two main groups:

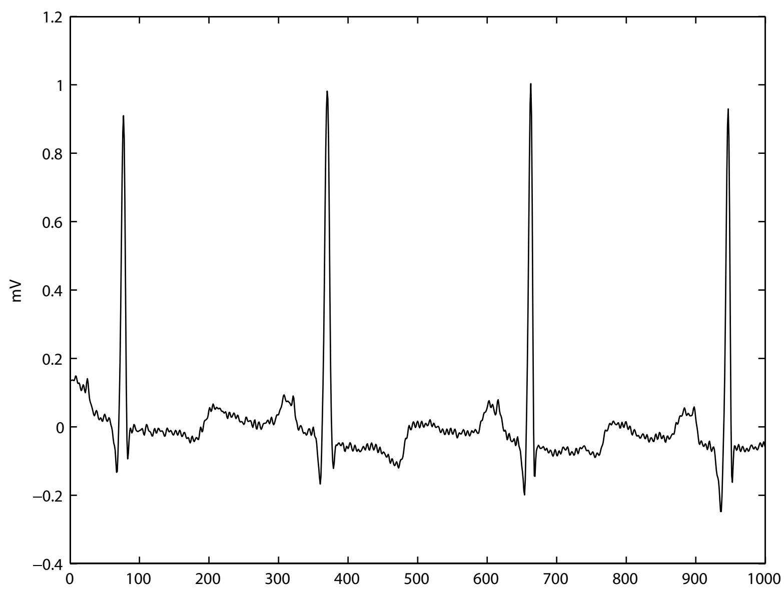
Figure 4. (a) ECG signal (top); (b) transformed coefficients (bottom)



1. Dedicated techniques:
   a. Direct Methods: Amplitude Zone Time Epoch Coding (AZTEC) (Cox, Noelle, Fozzard, & Oliver, 1968), Turning Point (TP) (Mueller, 1978), Coordinate Reduction Time Encoding System (Abenstein & Tompkins, 1982), FAN algorithm (Dipersio & Barr, 1985) and modified time domain coding algorithms, such as SLOPE (Tai, 1991) and Scan Along Polygonal Approximation (SAPA) (Shahein & Abbas, 1994).
   b. Optimization algorithms: Long-Term Prediction (LTP) (Nave & Cohen, 1993), algorithms based on analysis by synthesis (ASEC) (Jalaleddine, 1990), and the Cardinality Constrained technique (Nygaard & Hauglan, 1998).
2. Generic techniques: These techniques can be used with a great variety of signals, including voice, image, and video. This includes Differential Pulse Code Modulation (DPCM), Subband Coding (SC) and transform-based coding (Hilton, 1997; Istepanian, Hadjileontiadis, & Panas, 2002; Lu, Kim, & Pearlman, 2000; Rajoub, 2002).

Figure 5. (a) original ECG signal (top); (b) reconstructed ECG signal after discarding 80% of the coefficients (bottom)





Currently, the algorithms that provide the best ECG compression results use transform-based coding, and the most widely used is the wavelet transform. These algorithms can be grouped, with respect to coefficient encoding, into two main groups: algorithms with coefficient thresholding, and algorithms with tree-based coefficient encoding.

Coefficient thresholding algorithms follow a methodology that was presented in the previous section. The majority of these algorithms also apply lossless compression techniques to encode the wavelet coefficients. One example of this technique is the work by Rajoub (2002).

Tree-based coefficient encoding algorithms use hierarchic structures for encoding the wavelet coefficients, by exploiting the redundancy in the scales of the wavelet decomposition. Such algorithms provide results superior to traditional thresholding algorithms. Among recent works that use this paradigm, we can mention the works of Istepanian et al. (2002), Hilton (1997), and Lu et al. (2000).

Figure 6. Two-dimensional wavelet decomposition

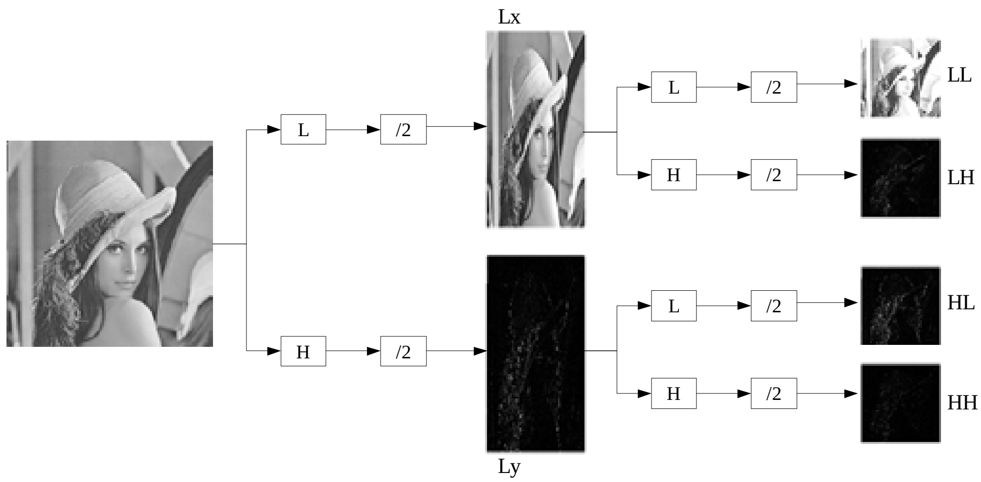
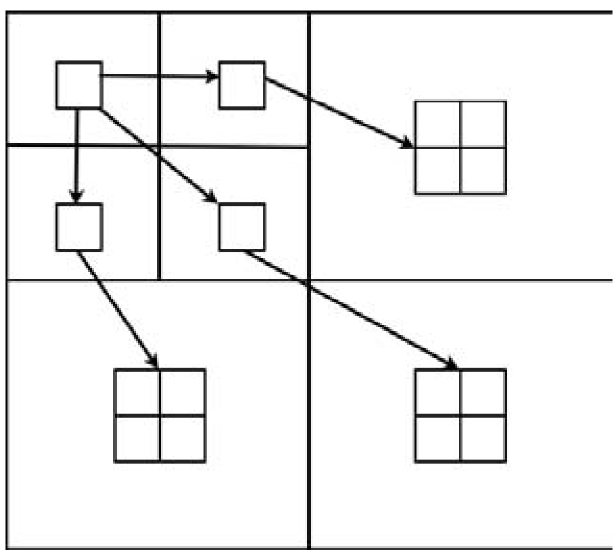


Figure 7. Wavelet coefficients tree



The main techniques described in the literature for EMG signal compression can be found in (Berger, Nascimento, Carmo, & Rocha, 2006; Guerrero & Maihes, 1997; Norris, 2001; Norris, Englehart, & Lovely, 2001; Wellig, Zhenlan, Semling, & Moschytz, 1998). Norris et al. (1995) investigates the compression of EMG signals using Adaptive Differential Pulse Code Modulation (ADPCM). Guerrero and Maihes (1997) compared different methods for lossless compression with methods based on orthogonal transforms. The orthogonal transform based showed better performance, with respect to compression ratio.

Recently, the use of wavelet transform for compression of EMG signals was studied by Wellig et al. (1998), Norris et al. (2001), and Berger et al. (2006). Wellig et al. (1998) and Norris et al. (2001) encoded the transformed coefficient using the EZW algorithm and compared their results with the traditional wavelet transform. Berger et al. (2006) used a scheme for adaptive dynamic bit allocation for encoding the coefficients. These methods provide performance equivalent or better than previously described methods.

Comparing different algorithms for EMG signal compression is difficult, because there is no standard database, such as the *MIT-BIH Arrhythmia Database*, which is used in most studies on compression of ECG signals. The studies on EMG signal compression usually use different test signals, involving different muscles, volume conductors, conduction velocities, loads, sampling rates and duration. Therefore, the comparison between different algorithms should always be analyzed with caution.

The algorithms mentioned in this article have the goal of guaranteeing the fidelity of the reconstructed signal after decoding. In this sense, the wavelet transform and the EZW and SPIHT algorithms are considered by the specialized public as the most promising lines of work, and future research tend to emphasize improvements in these algorithms.

However, reaching a good quality representation of signals with low bit rates comes with the cost of increasing the complexity of the compression algorithms. Few studies on low-cost real-time compression of biomedical signals have been published, even though the interest on telemedicine applications is growing. This is an important issue, which should be addressed in the future. The works by Alesanco, Olmos, Istepanian, and García (2003) and Kim, Yoo, and Lee (2006) proposed the first solutions for real-time ECG compression, and Carotti, De Martin, Merletti, and Farina (2006) a lossy coding technique for surface EMG signals that is based on the algebraic code excited linear prediction (ACELP) paradigm, widely used for speech signal coding. The algorithm was adapted to the EMG characteristics and tested on both simulated and experimental signals. This method is characterized by moderate complexity (approximately 20 million instructions/s) and an algorithmic delay smaller than 160 samples (~160 ms).

CONCLUSION

This article presented a review of techniques for compression of biomedical signals. Among the different techniques that were presented, the ones based on wavelet transforms had the best performance with respect to compression gain and reconstructed waveform fidelity. When the wavelet transform is used, a great challenge is to find effective alternatives for efficiently encoding the transformed coefficients. The most efficient methods for this are the EZW and SPIHT algorithms. An important line of future work is the search for real-time, low-consumption algorithms, which would be suitable for telemedicine applications.

REFERENCES

Abenstein, J., & Tompkins, W. (1982). New data-reduction algorithm for real-time ECG analysis. *IEEE Trans. Biomed. Engineering. 29*, 43–48.

Alesanco, A., Olmos, S., Istepanian, R., & Garcia, J. (2003). A novel real-time nultilead ECG compression and de-noising method based on the wavelet transform. *Computers in Cardiology*, 30, 593-596.

Alesanco, A., Olmos, S., Istepanian, R.S.H., & García, J. (2006) Enhanced real-time ECG coder for packetized telecardiology applications. *IEEE Transactions on Information Technology in Biomedicine, 10*(2), 229–236.

Berger, P.A., Nascimento, F.A.O., Carmo, J.C., & Rocha, A.F. (2006) Compression of EMG signals with wavelet transform and artificial neural networks. *Physiological Measurement, 27*, 457–465.

Carotti, E., De Martin, J.C., Merletti, R., & Farina, D. (2006). Compression of surface EMG signals with algebraic code excited linear prediction. *Medical Engineering & Physic, 29* (2), 253–258.

Cox, J., Noelle, F., Fozzard, H., & Oliver, G. (1968). AZTEC: A preprocessing program for real-time ECG rhythm analysis. *IEEE Trans. Biomed. Engineering. 15*, 128–129.

Daubechies, I. (1988). Orthogonal bases of compactly supported wavelets. *Comm. Pur. Appl. Math. 41*, 909–996.

Dipersio, D.A., & Barr, R.C. (1985). Evaluation of the FAN method of adaptive sampling on human eletrocardiograms. *Med. Biomed. Eng. Computing*, 401–410.

Guerrero, A., & Maihes, C. (1997). On the choice of an electromyogram data compression method. In *Proceedings of the 19th annual International Conference of the IEEE Engineering in Medicine Biology Society* (pp. 1558–1561).

Gersho, A., & Gray, R. (1992). *Vector quantization and signal compression*. Kluwer Acad. Publication.

Hilton, M.L. (1997). Wavelet and wavelet packet compression of electrocardiograms. *IEEE Trans. Biomed. Engineering, 44*, 394–402.

Huffman, D.A. (1952). A method for the construction of minimum redundancy codes. In *Proceedings of IRE* (Vol. 40, No. 10, pp. 1098–1101).

Istepanian, R.S.H., Hadjileontiadis, L., & Panas, S. (2001). ECG data compression using wavelets and higher order statistics methods for telemedical applications. *IEEE Transactions on Information Technology in Biomedicine, 5*(2), 108–115.

Jalaleddine, S.M., Hutchens, C.G., Satrattan, R.D., & Coberly, W.A. (1990). ECG data compression techniques—The unified approach. *IEEE Trans. Biomed. Engineering, 37*(4), 329-343.

Jayant, N.S., & Noll, P. (1982). *Digital coding of waveforms – principles and application to speech and video*. Prentice Hall.

Ken, C. P. (1985). *Principles of digital audio*. Sams/Prentice-Hall Computer Publishing.

Kim, S.B., Yoo, S.K., & Lee, M.H. (2006). Wavelet-based low-delay ECG compression algorithm for continuous ECG transmission. *IEEE Transactions on Information Technology in Biomedicine, 10*(1), 77–83.

Lu, Z., Kim, D.Y., & Pearlman, W.A. (2000). Wavelet compression of ECG signals by the set partitioning in hierarchical trees (SPIHT) algorithm. *IEEE Trans. Signal Processing, 47*, 849–856.

Mallat, S. (1989). A theory for multiresolution signal decomposition: The wavelet representation. In *Proceedings of IEEE Trans on Pattern Anal. and Math. intel.* (Vol. 11, No. 7).

Mueller, W. (1978). Arrhythmia detection program for an ambulatory ECG monitor. *IEEE Trans. Biomed. Engineering, 14*, 81-85.

Nave, G., & Cohen, A. (1993). ECG compression using long-term prediction. *IEEE Trans. Biomed. Engineering, 40*, 877-885.

Norris, J.A., Englehart, K., & Lovely, D. (2001). Steady-state and dynamic myoelectric signal compression using embedded zero-tree wavelets. In *Proceedings of the 23th annual International Conference of the IEEE Engineering in Medicine Biology Society* (pp. 1879–1882).

Nygaard, R., & Haugland, D. (1998). Compressing ECG signals by piecewise polynomial approximation. In *Proceedings of the Int. Conf. Acoust. Speech Signal Processing* (pp. 420-424). Seattle, WA.

Rajoub, B.A. (2002). An efficient coding algorithm for the compression of ECG signals using the wavelet transform. *IEEE Trans. Biomed. Eng., 49*(4), 355–362.

Said, A., & Pearlman, W.A. (1996). A new fast and efficient image codec based on set partitioning in hierarchical Trees. *IEEE Transactions on Circuits and Systems for Video Technology, 6*, 243–250.

Sayood, K. (1996). Introduction to data compression. In C.F. Shahein, & H.M. Abbas (Eds.), ECG Data-Compression via Cubic-Splines and Scan Along Polygonal Approximation. *Signal Processing, 35*(3), 269–283. Morgan Kaufman.

Shahein, C.F., & Abbas, H.M. (1994). ECG data-compression via cubic-splines and scan along polygonal approximation. *Signal Processing, 35*(3), 269-283.

Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Tech. Journal, 27*, 379-423.

Sörnmo, L., & Laguna, P. (2006). Electrocardiogram (ECG) signal processing. In M. Akay (Ed.), *Wiley Encyclopedia of Biomedical Engineering* (Vol. 2, pp. 1298–1313). Wiley.

Tai, S.C. (1991). Slope – a real-time ECG data compressor. *L Biomed. Eng. Computing, 29*, 175–179.

Wellig, P., Zhenlan, C., Semling, M., & Moschytz, G.S. (1998). Electromyogram data compression using single-tree and modified zero-tree wavelet encoding. In *Proceedings of Engineering in Medicine and Biology Society 23rd Annual International Conference of the IEEE* (pp. 1303–1306).

Witten, I., Neal, R., & Cleary, J. (1987). Arithmetic coding for data compression. *Communications of the ACM, 30*(6).

B

KEY TERMS

Arithmetic Coding: A more sophisticated compression technique, based on Huffman coding concepts. It results in compression gains closer to the theoretical limits. In this coding, character sequences are represented by individual codes depending on their probability of occurrence.

Biomedical Signals: Refers to signals that carry useful information for probing, exploring, and understanding the behavior of biomedical systems under investigation.

Delta Coding: Refers to signal compression techniques that store a digital signal as the difference between successive samples. The first sample in the encoded signal is equal to the first sample of the original signal. Subsequent encoded samples are equal to the difference between the current sample and the previous sample of the original signal.

Direct Methods: Encode signals in time domain. These algorithms depend on the morphology of the input signal. Direct compression methods use sophisticated processes and "intelligent" signal decimation and interpolation.

Discrete Cosine Transform: A technique for expressing a waveform as a weighted sum of cosines. The DCT is central to many kinds of signal processing, especially video compression.

Embedded Zerotree Wavelet: A simple, yet remarkable effective, compression algorithm witch exploits the self-similarity of the wavelet transform coefficients at different scales, by structuring the data in hierarchical tree sets that are likely to be highly correlated, yielding a fully embedded code.

Huffman Coding: Refers to a compression technique where the data are represented as a set of variable length binary words. The lengths depend on the frequency of occurrence of the symbols used for representing each signal value. Characters that are used often are represented with fewer bits, and those that are seldom used are represented with more bits.

Lossless Compression: A signal compression technique where the original signal can be perfectly recovered from the compressed signal. The Morse code of telegraphy provides a classic example: by sending short patterns (dots and dashes) for frequent letters, and long patterns for rare letters, fewer bits are needed on the average than with the standard computer representation (ASCII), which assigns the same number of bits to all letters.

Lossy Compression: Compression is lossy if the original signal cannot be perfectly recovered from the compressed file, as a controlled amount of distortion is allowed. Lossy signal compression techniques show higher compression gains than lossless ones.

Run-Length Coding: A lossless compression. A set of symbols is represented as a sequence of RUN/LEVEL. RUN being the number of consecutive zeros and LEVEL being the value of the following nonzero coefficient.

Set Partitioning in Hierarchical Tree: A refinement of the Embedded Zerotree Wavelet algorithm. SPIHT requires partitioning the wavelet coefficients into a number of lists, with list membership changing as the execution proceeds, and in some cases involving dual membership of a coefficient in different lists.

Signal Compression: The reduction of the necessary amount of bits to represent a signal. In many cases, this must be accomplished while preserving the morphological characteristics of the waveform. In theory, signal compression is the process where the redundant information contained in the signal is detected and discarded.

Transform Compression: Compression technique that is based on a simple premise: when the signal is processed by a transform, the signal energy (information) that was distributed among all the samples in the time domain can be represented with a small number of transformed coefficients.

Wavelet Transform: Refers to a transformation where a signal containing N samples is first filtered by a pair of filters that decompose the signal into low and a high frequency bands. Each band is undersampled by a factor of two, that is, each band contains N/2 samples. With the appropriate filter design, this action is reversible.

Biosensors

Sourabh Bansal
Illinois Institute of Technology, USA

INTRODUCTION

Biosensor[1] is a diagnostic tool in which a biological element is used to sense a chemical and its amount in a given sample, and then the sensed information (i.e., data) is transferred to a transducer which converts this signal to electrical signal. In this way, it transforms a biological response into an electrical signal. It also detects, records, and transmits data generated due to physiological change or any chemicals presence in the area being analyzed. The analysis is accurate and reliable.

In other words, biosensor can be termed as a device used in a biological derived sensing element[2] integrated with a physiochemical transducer, producing an electrical signal (Turner, 1996). The resulting electrical signal is a measure of the amount of chemical or combination of chemicals being detected.

Sometimes Biosensors are referred as the living organisms, which are used as a sensor to detect the environmental change.

BACKGROUND[2]

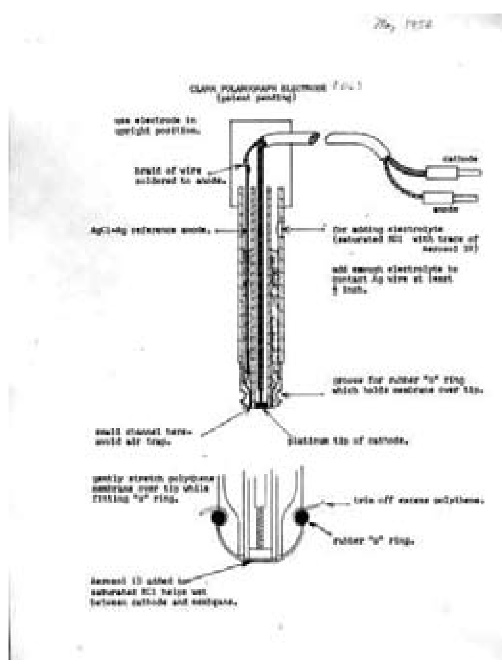
Professor Leland C. Clark Jr. led the foundation of biosensor concept. He was a pioneer in this field. His name will always remain embedded with the concept of *Artificial Blood*[3]. In 1956, Clark ignited the biosensor concept by publishing one of his revolutionary papers on the oxygen electrode (Figure 1).

Soon, Clark invented a device to rapidly determine the amount of glucose present in the blood at a given instant of time. In 1962, at the New York Academy of Sciences symposium, based on his experience and addressing his desire to expand the range of analytes that could be measured in the body, he made an ineffaceable mark by describing how to make electrochemical sensors more intelligent by adding enzyme transducers as membrane enclosed sandwiches. After that, Guilbault and Montalvo described a potentiometric enzyme electrode to sense urea, based on urease immobilized at an ammonium-selective liquid membrane electrode. In 1974, thermal transducers for biosensors were proposed, and were named as thermal enzyme probes and enzyme thermistors respectively.

Soon, Clark's spark was turned into fire in 1975 with the successful launch of Yellow Springs Instruments Company glucose analyzer, based on the amperometric detection of hydrogen peroxide. In the same year, a paper regarding utilizing bacteria as the biological element in microbial electrodes for the measurement of alcohol was presented by Divis. Later, during the same year, Lubbers and Opitz introduced the term "optode" to describe a fiber-optic sensor with immobilized indicator to measure Carbon dioxide or oxygen. Commercial optodes are very well used in vivo measurements of pH, pCO_2 and pO_2. After that, a lot of ideas were floating, but as the time passed, they sank.

In 1982, Shichiri was the first one to describe the first needle-type enzyme electrode for subcutaneous implantation. In 1984, Cranfield Biotechnology Centre, from Cranfield University, published a paper on the use

Figure 1. The Clark polarograph electrode



of ferrocene and its derivatives as an immobilized mediator for use with oxidoreductases in the construction of inexpensive enzyme electrodes. In 1987, based on this, MediSense (Cambridge, USA) launched screen-printed enzyme electrodes with a pen-sized mater for home blood-glucose monitoring. Today, 18.2 millions of diabetic people in USA[4] depend on Clark's original glucose sensor concept for self-monitoring.

BASIC STRUCTURE OF BIOSENSOR[5]

Biosensor comprises of a biological element which is connected to a transducer, which in turn is connected to electronics, which can process the received information, and then finally display it. The illustrated version has been demonstrated in Figure 2.

A most common example of a biosensor for the detection of glucose in a sample (blood) can be seen in Figure 3.

Glucose oxidase enzyme (GOx) in glucose molecule is converted to hydrogen peroxide (H_2O_2) and gluconic acid ($C_6H_{12}O_7$) within the biolayer. The H_2O_2 passes through all the layers in the path, and is detected at platinum electrode. The electrode senses it, and the current is thus produced, which is then processed further, and is passed on to the display unit to show the concentration of glucose in given sample.

CLASSIFICATION OF BIOSENSORS[5]

Broadly biosensors can be classified according to Transducers, bioactive components, or immobilization techniques used in the biosensor. Detailed classification of Biosensors has been illustrated in Figure 4, 5 and 6.

Transducer used can be Electromechanical, optical, piezo-electrical, or thermal. Bioactive components can be enzymes, antibodies, cells, nucleic acids, or lipids. Immobilization techniques can be physical or chemical.

CURRENT AND FUTURE DEVELOPMENTS IN BIOSENSORS

Biosensor is the one of the most rapidly growing fields now-a-days. Its current developments and focus is mainly towards the health care area. In spite of its tremendous potential, many of the developments in Biosensors are still not commercialized. As of now, mostly single analyte devices are present; however, in the future, multianalyte sensor will be there. Research and development in biosensor field is very rapid, and lot of potential is there in coming years:

- Routine analytical measurement of folic acid, biotin, vitamin B12, and pantothenic acid[6].
- Remote sensing of airborne bacteria (e.g., in counter-bioterrorist activities[6]).
- Most of the research is going on in the development of noninvasive devices strips to diagnose diseases[7].
- Medical telesensor is in development phase, consisting of an array of chips to collectively monitor bodily function such as blood pressure, oxygen level, pulse rate, and body temperature[8].

Figure 2. Basic structure of biosensor (Source: http://intel.ucc.ie/sensors/Biosens.htm)

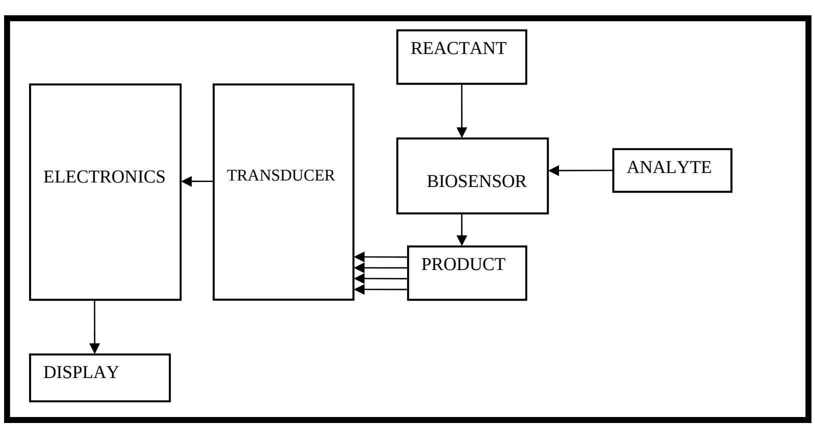


Figure 3. Biosensor for detection of glucose (Source: http://intel.ucc.ie/sensors/animation.htm)

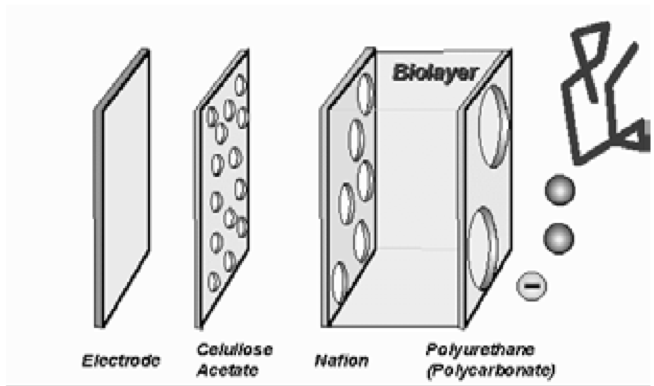


Figure 4. Classification of biosensor based on transducer used (Source: http://intel.ucc.ie/sensors/Biosens.htm)

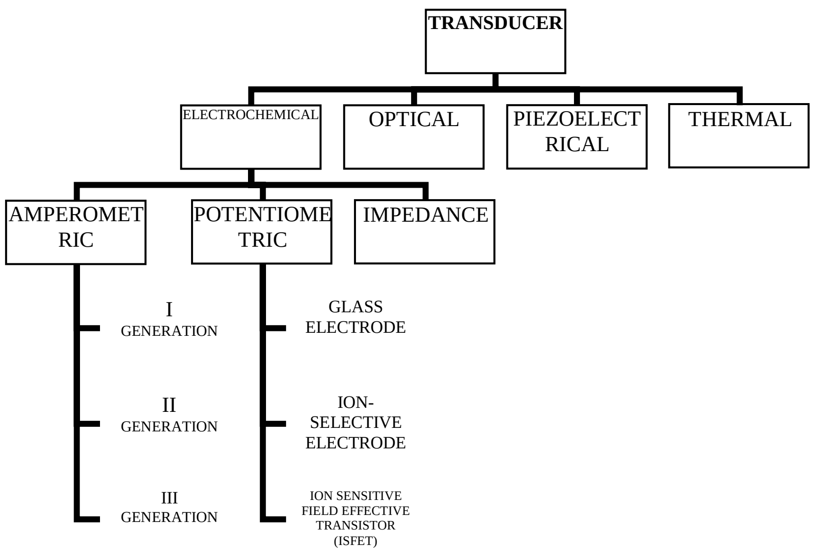


Figure 5. Classification of biosensor based on bioactive components used (Source: http://intel.ucc.ie/sensors/Biosens.htm)

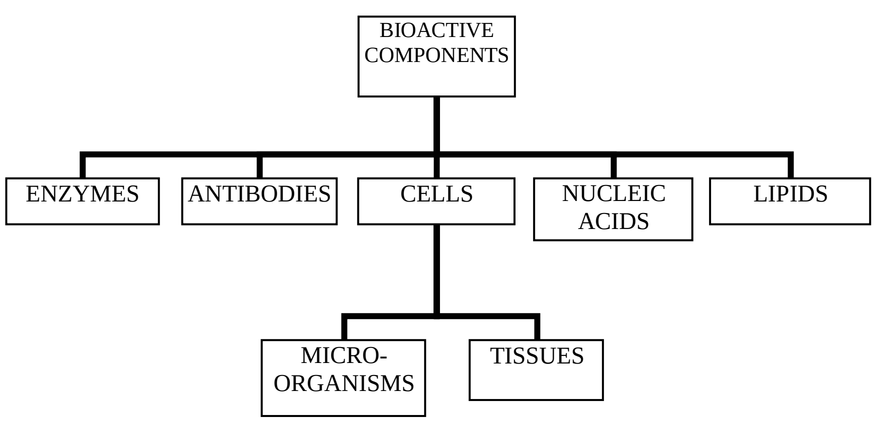
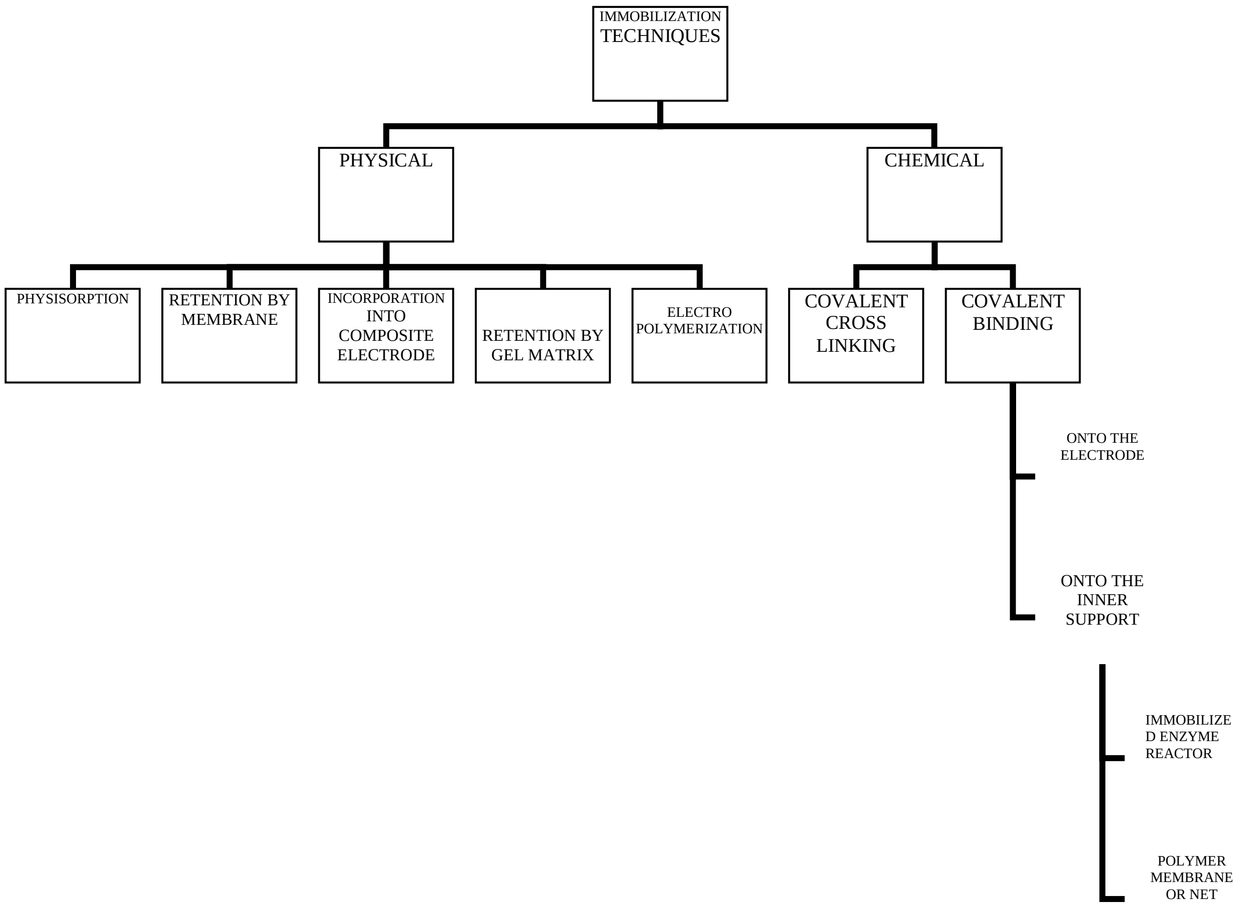


Figure 6. Classification of biosensor based on immobilization technique used (Source: http://intel.ucc.ie/sensors/Biosens.htm)



- In military, biosensors are being pursued to detect and analyze pathogenic microorganisms or any contaminates in food[9].
- Research is also being done in the field of environment monitoring. A biosensor with very high sensitivity and designed for real time analysis for contaminates (arsenic, lead, pesticides, phenol, and so on) in ambience is under development. This kind of device helps in field screening at given instant of time[10].
- ORNL (Oak Ridge National Laboratory) has been developing a lot of biosensors using sophisticated technology to detect a specific abnormality in living beings. The major one is the optical biopsy sensor developed by Vo-Dinh with researchers at Thompson Cancer Survival Center in Knoxville (Vo-Dinh, Kasili, & Wabuyele , 2006). This laser-based fluorescence sensor can depict whether a tumor is a cancerous or not. Its discovery has made the determination of cancer of esophagus easier by elimination of biopsy, reducing pain in patients[8]

CONCLUSION

Biosensors will bring a drastic revolution in health care and medicines. The Biosensor field is rapidly being developed. Biosensors have a lot of potential in the future, whether it is the area of defense, industrial, medical, environmental, or food related areas. Glucose monitoring has proved a great help for the people who are suffering from diabetes, or who are on the verge of it. A lot of research is going on in different areas around the globe.

B

REFERENCES

Chen, L., McBranch, D.W., Wang, H.L., Helgeson, R., Wudl, F., & Whitten, D.G. (1999). Highly sensitive biological and chemical sensors based on reversible fluorescence quenching in a conjugated polymer. In *Proceedings of Natl Acad Sci U S A* (Vol. 96, No. 22, pp. 12219–12221).

Clark, L.C. Jr. (1993). How the first enzyme electrode was invented. *Biosens Bioelectron*, *8*(1), iii–vii.

Cooper, M.A. (2006). Optical biosensors: Where next and how soon? *Drug Discov Today*, *11*(23 –24), 1061–1067.

Deisingh, A.K. (2006). Aptamer-based biosensors: Biomedical applications. *Handb Exp Pharmacol.*, *173*, 341–57.

Ekinci, K.L. (2005). Electromechanical transducers at the nanoscale: Actuation and sensing of motion in nanoelectromechanical systems (NEMS). *Small*, 1(8–9), 786–797.

Golightly, G. (2000). Biosensors May Revolutionize Space Life Support. *Ball Semiconductor Inc.* Retrieved February 11, 208, from http://www.space.com/businesstechnology/technology/biosensors_life_000125.html

Hansen, K.M., & Thundat, T. (2005). Microcantilever biosensors. *Methods*, *37*(1), 57–64

Harms, H., Wells, M.C., & van der Meer, J.R. (2006). Whole-cell living biosensors--are they ready for environmental application? *Appl Microbiol Biotechnol*, *70*(3), 273–280.

Heeger, P., & Heeger, A. (1999). Making sense of polymer-based biosensors. In *Proceedings of Natl Acad Sci U S A.* (Vol. 96, No. 22, pp. 12219–12221).

Jacobson, B.K. Biosensors and other medical and environmental probes. Retrieved February 11, 2008, from http://www.ornl.gov/info/ornlreview/rev29_3/text/biosens.htm

Leone, B.J. (1998). Artificial blood: What is it? Will I use it? Retrieved February 11, 2008, from http://www.dcmsonline.org/jax-medicine/1998journals/december98/artificialblood.htm

Rasooly, A., & Herold, K.E. (1989). Biosensors for the analysis of food- and waterborne pathogens and their toxins. *J AOAC Int.*, 89(3), 873–883.

Teicher, B. (1993). Use of perfluorocarbon emulsions in cancer therapy. *Blood Substitutes and Oxygen Carriers*, 640–647.

Turner, A.P.F. (1996). Biosensors: Past, present and future. Retrieved February 11, 2008, from http://www.cranfield.ac.uk/biotech/chinap.htm

Vo-Dinh, T., Kasili, P., & Wabuyele, M. (2006). Nanoprobes and nanobiosensors for monitoring and imaging individual living cells. *Nanomedicine*, 2(1), 22–30.

KEY TERMS

Amperometric: Pertaining to or involving the measurement of an electric current.

Electrochemical: Of or relating to chemical reactions brought about by electricity; galvanochemical.

Electrode: A conductor, not necessarily metallic, through which a current enters or leaves a nonmetallic medium, as an electrolytic cell, arc generator, vacuum tube, or gaseous discharge tube.

Oxidoreductases: An enzyme that catalyzes an oxidation-reduction reaction.

Piezoelectrical: Of or relating to generation of electricity, or of electric polarity in dielectric crystals subjected to mechanical stress.

Potentiometric: Pertaining to or involving the measurement of an electromotive force, or potential difference by comparison with a known voltage.

Subcutaneous: Performed or introduced under the skin, as an injection by a syringe.

ENDNOTES

[1] http://en.wikipedia.org/wiki/Biosensor

[2] http://www.cranfield.ac.uk/biotech/chinap.htm

[3] http://www.dcmsonline.org/jax-medicine/1998journals/december98/artificialblood.htm

[4] http://medgadget.com/archives/2005/02/dr_leland_c_cla.html
[5] http://intel.ucc.ie/sensors/Biosens.htm
[6] http://en.wikipedia.org/wiki/Biosensor
[7] http://www.biodesign.asu.edu/centers/bb/projects/
[8] http://www.ornl.gov/info/ornlreview/rev29_3/text/biosens.htm
[9] http://www.natick.army.mil/Soldier/media/fact/food/biosensor.htm
[10] http://www.biodesign.asu.edu/centers/bb/projects/

B

Blood Pressure Estimation with Considering of Stroke Volume Effect

Moha'med O. Al-Jaafreh
University of Technology, Australia

Adel A. Al-Jumaily
University of Technology, Australia

INTRODUCTION

The mean arterial pressure (MAP) is a very important cardiovascular parameter for physicians to diagnose various cardiovascular diseases. Many algorithms were used to estimate MAP with different accuracy. These algorithms used different factors, such as blood level, pulses, and external applied pressure, photo-plethysmography (PPG) signal features, heart rate (HR), and other factors. In addition, some natural-based techniques were employed to minimize the difference between estimated and measured blood pressure, as well as to measure blood pressure continuously.

This article presents an algorithm to estimate MAP, utilizing the HR, Stroke Volume (SV), and Total Peripheral Resistance (TPR), with considering SV changing influence; this consideration is investigated mathematically, and by the Particle Swarm Optimization (PSO) technique.

This new algorithm was implemented on 20 cases of mimic database (Mimic Database, http://www.physionet.org/physiobank/database/mimicdb); 10 cases for training to investigate the suitable values of these factors, and all cases for verification and examining the performance of these techniques. The achieved results of estimated MAP by these techniques are compared with real MAP, and illustrated in tables and figures.

BACKGROUND

The cardiovascular system consists of heart, vessels, and blood. In a healthy person, the heart pumps the blood in vessels with synchronous pulses (HR) and pulse wave velocity (PWV). The force of blood flux, which is caused by heart beating, forms a pressure against blood vessels' walls.

Blood pressure (BP) is a vital measurement used by the physicians for diagnosing the health situation of subjects, and saving them from critical diseases or some dangerous circumstances, such as hypertension, hypotension, artery stiffness, coma or heart attack.

Many models have used to estimate BP; these models rely on factors such as blood level, pulses, external applied pressure, PPG signal features, HR, artery volume changes, and many other factors.

The first model has been used to estimate BP, and depends on the blood level factor (O'Brien & O'Malley, 1981); blood level is related to BP, according to Equation 1:

$$P = \rho \times g \times h \qquad (1)$$

Where P is blood pressure,
ρ is blood density,
g is acceleration of gravity,
h is blood level.

The blood level is measured by inserting a catheter on the subject's artery; the blood level is elevated in the inserted catheter by BP affect, and this blood pressure is computed by Equation 1 and displayed on a monitor or drawn on paper. This method is known as the arterial cannulation method—it was first used by Ludwig on 1847. It is an accurate and reliable method, but it is very invasive, and causes limb ischemia, thrombosis, hemorrhages, and other bad affects (Kasirajan, Simmons, King, Shumaker, DeAnda, & Higgins, 2002).

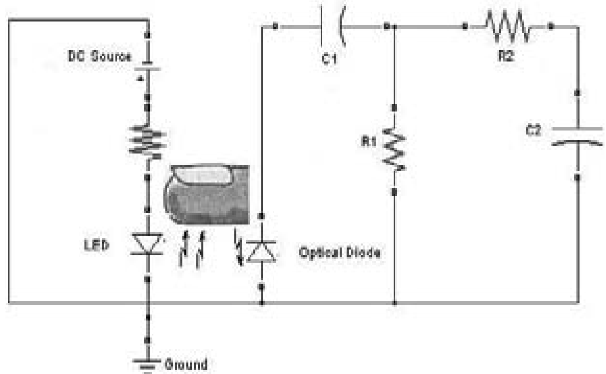
The second method to estimate BP is the Auscultatory method. It is first used by Korotkoff in 1905 (Beecher, 2003). This method depends on heart pulses and external applied pressure factors. First, the blood flow is obscured by applied external pressure on an artery by inflation of an air cuff; second, the pressure is reduced slowly. Finally, the blood pressure equals applied pressure when the blood starts flowing on the squeezed artery, and the heart pulse is heard by an expert physician or nurse, while reducing that applied pressure (Geddes, Hoff, & Badger, 1966).

This method has to be implemented by a physician or nurse by using sphygmomanometer, which consists of a cuff to apply pressure, and a mercury (Hg) meter to measure that pressure. The Auscultatory method is more widely used than the arterial cannulation method, because it is not invasive. But it is a discontinuous method, because of the inflation and defilation of the cuff, which is wrapped around the subject's arm. Also, it is uncomfortable, because of the affect of applied pressure.

On other hand, BP has been estimated depending on pulse arrival time and PWV (Bazett & Dreyer, 1922; Lansdown, 1957). Afterwards, Mackay (1964) estimated BP by piezoelectric pressure transducers, and developed the tonometry method. Soon after, depending on PPG and artery volume changes, some methods were developed to estimate BP, such as volume-compensation and local pressurization. In addition, pulse wave velocity and pulse arrival time have been used to estimate BP changes (Chen, Kobayashi, Ichikawa, Takeuchi, & Togawa, 2000; Geddes, Voelz, Babbs, Bourland, & Tacker, 1981; Piston & Stradling, 1998).

Finally, the cuffless and real-time measurement methods have been developed to estimate BP. Actually, these methods are noninvasive, continuous, and comfortable methods, because they relied on a PPG signal. The PPG sensor continuously emits light by a light-emitting diode, as a transmitter, through the subject's finger, and then the PPG signal is extracted by an optical diode as a receiver. A PPG sensor circuit is shown in Figure 1.

Figure 1. Circuit model



The cuffless and real-time measurement methods are used to estimate BP, by using the PPG signal's features through mathematical models, which depend on the relationship between BP and the PPG signal's features. The BP and PPG signal's features' relationship is complex and dynamic (Shaltis, Reisner, & Asada, 2005). This relation has been studied—Yan and Zhang (2005) stated that the normalized harmonic area of a discrete period transform of the PPG signal has a high correlation with BP. Moreover, the time interval between the peak of electrocardiograph (ECG) and the front foot of the PPG signal, also known as the PTT_f; has a high correlation with BP variability (Ma & Zhang, 2005).

Furthermore, a model has been found to estimate BP by using a weighted PTT (PTT_w); this feature is dependent on that specified subject. PTT_w depends on calibrating measurements of systolic and diastolic BP (Poon & Zhang, 2005).

Moreover, some work of PPG-based methods has established correlation between the frequency of the PPG signal, HR, and BP (Ganong, 2003). MAP is related to HR, as shown in Equation 2:

$$MAP = HR \times SV \times TPR \qquad (2)$$

Where MAP is mean arterial pressure,
SV is stroke volume,
HR is heart rate,
TPR is total peripheral resistance.

This correlation depends on constant factors SV and TPR. In reality, these factors are not constant, and the relationship between MAP and HR is nonlinear.

In our work, a new algorithm is developed to estimate MAP, utilizing the HR, stroke volume, and total peripheral resistance, and considering SV changing influence; this consideration is investigated mathematically, and by the Particle Swarm Optimization technique.

This new algorithm was implemented on 20 cases of mimic database (refer to Mimic Database Web site); 10 cases for training to investigate the suitable values of these factors, and all cases for verification, and examining the performance of these techniques.

METHODS TO ESTIMATE MAP

Practically, the relationship between MAP and HR is nonlinear (Shaltis et al., 2005) because of many reasons, such as SV changing, and TPR. Therefore, PSO is used to deal with this nonlinear research problem.

A new mathematical model was developed, considering SV changing, which depends on HR. Therefore, Equation 2 was written as:

MAP = HR × SV × TPR ---------- (2)

MAP = HR × (SV_0 + A × SV_0 ×HR) × TPR

MAP = SV_0 ×TPR ×HR× (A × HR+1) ------- (3)

where SV_0 is the constant stroke volume of the heart.
A is the coefficient of SV changing.

So how can A—coefficient of changing—be calculated? In this article, A is calculated mathematically, and by using PSO, by considering twenty cases of MIMIC database. The HR and MAP have been extracted from 20 cases of MIMIC database to investigate the relationship between them.

MIMIC database is available in Physio-bank database (refer to Mimic Database Web site), which includes continuous records for each case from the intensive care unit (ICU) every second over 10 hours; that means 36,000 records for each case, at least. In addition, these records include signals and interrupted measurements such as HR, MAP, Systolic BP, and Diastolic BP for each case, which are measured from a bedside monitor and clinical database extracted from patients' medical records.

First, the following procedure has been followed to calculate A mathematically by fit function:

1. The coefficient of SV changing percentage (A) was generated by using a fit function for 10 cases processed by developed formula, which is:

   MAP = SV_0 ×TPR ×HR× (A × HR+1) ------- (3)

2. The best A is when the mean of error for all cases is the smallest.

   where Error = Real MAP – Estimated MAP

3. The MAP was estimated for all cases' records by equation (3):

   Where TPR = 0.018 mmHg.min/ml (Klabunde, 2006)
   SV_0 = 70 ml (Klabunde, 2006).
   A = Achieved value of A by fit function

Second, A is calculated by the PSO technique—PSO is a new based nature technique, and has been invited by Russell Eberhart and Jim Kennedy in 1996. PSO deals with various types of nonlinear continuous problems. PSO is utilized to find an optimal solution for nonlinear continuous problems by simulating the social model and imitating the birds and other animals' movement within swarm topologies.

In general, PSO ties to artificial life, while in particular, it ties to bird flocking, fish schooling, and swarming theory. Also, PSO related to evolutionary computation, and had ties to both genetic algorithms and evolutionary programming (Kennedy, Eberhart, & Shi, 2001).

Particle Swarm Optimization fundamentals are the population of particles, interconnection topologies, search algorithms, and evaluation rules; these fundamentals cooperate to find the optimum solution of the problem.

The normal population of PSO is 20 to 50 particles, this number is determined depending on the problem size; however, this population is far less than usual for population of other evolutionary algorithms.

PSO particles have interconnection topologies describing the communication among them to move within problem domains to find optimum solutions. There are two topologies:

1. Sociometry topology, or global best topology, where every particle is connected with all particles of a population, and influenced by a particle which has found the best problem solution.
2. Ring lattice, or local best topology, where every particle is connected to previous and the next particles of the population array only.

Local best topology can converge separately on various optima solutions in problem space, and it has fewer connections between particles than global best topology. However, global best topology is faster to find an optimum solution of problems (Zomaya, 2006).

The main fundamental of PSO is the search algorithms, which change particles' positions through these equations:

$$xd\ (t + 1) = xd\ (t) + vd\ (t + 1) \qquad (4)$$

$$vd(t + 1) = \chi \times vd(t)\ \varphi \times rand \times (pi,d - xi,d(t)) + \varphi 2 \times rand \times (pg,d - xi,d(t)) \qquad (5)$$

v_d is the displacement of the particle's movement
x_d is the particle's position within the problem domain
χ is the constriction coefficient
$\varphi_{1,2}$ is the acceleration constants
$p_{i,d}$ is the local best within neighbor particles
$p_{g,d}$ is the global best within all particles

First, parameters' values of $v_d(1)$, $x_d(1)$, $\varphi_{1,2}$, χ, $p_{i,d,}$ and $p_{g,d}$ are initialized to start particles' movements within the problem space.

Acceleration constants' values $\varphi_{1,2}$ manage particles' movements, and the probability of finding optimum solution slowly with high fineness, or quickly with less fineness. On the other hand, construction coefficient χ balances between the effect of previous displacement, and the effect of interconnection topologies local and global best on current displacement.

Besides these algorithms and their parameters, maximum displacement (V_{max}) and maximum position (X_{max}) will be located to surround the positions' movement within the problem space, and stop particle searching out of the problem space.

Second, particles fly within the problem domain by updating their displacements v_d by Equation 4, then adjusting their positions x_d by Equation 5.

Third, the evaluation rules appraise the obtained solution success, and assist PSO search algorithms to achieve the optimum problem solution by adjusting $p_{g,d}$ and $p_{i,d}$ values according to new particles' positions.

Finally, this operation is reiterated till the search algorithm reached the optimum solution.

In summary, PSO search algorithms and interconnection topologies control the particles' movements within the problem domain to find the optimum problem solution, which is evaluated by evaluation rules.

IMPLEMENTING PSO

The PSO is implemented to find a good estimation for MAP, through calculating A by the following procedure:

1. The problem domain for the PSO is the database of 10 cases of MIMIC database.
2. The swarm population is 30 particles.
3. The two particles' topologies—gbest and pbest—have been employed to achieve the solution.
4. Set constant values for swarm parameters (χ, $\varphi_{1,2}$, X_{max}, V_{max}) to move the particles within the problem domain to find the best suitable value for the coefficient of SV changing (A) to minimize the difference between real MAP and estimated MAP.
5. Then, testing all 20 cases' database to ensure that this is the optimal value for A, which minimizes the difference between real MAP and estimated MAP by Equation 3 for all cases.

For comparison, the estimated MAP is compared with real MAP and the estimated MAP by Equation 2:

MAP = HR × SV × TPR ---------- (2)

Where TPR = 0.018 mmHg.min/ml. SV = 70 ml.

These results and comparison between all MAP estimated by three methods, and real MAP, are illustrated in tables and figures, as shown in the next section.

RESULTS

Each case of MIMIC database has 36,000 records, at least. To illustrate these results properly, MAP for all records of all cases is estimated by three methods:

1. Equation 3 by substituting A of Fit function (Method 1).
2. Equation 3 by substituting A of PSO (Method 2).
3. Equation 2 (Method 3).

Then the means of estimated MAP (mmHg), absolute difference |real MAP-estimated MAP| (AD) and absolute percentage difference (APD) for all subjects of all methods were calculated. Where APD

$$= \left| \frac{\text{Real MAP - Estimated MAP}}{\text{Real MAP}} \right| \times 100\% .$$

Moreover, the averages of means of estimated MAP (mmHg), AD, and APD for all cases of all methods are computed, and then all are tabled in Table 1.

From Table 1, the average of mean of estimated MAP for all cases by Method 2 estimated MAP by Equation 3 by substituting A of PSO, also the average of AD and APD between mean of real MAP and mean of estimated MAP for all cases by Method 2 are the best and closest values to the mean of real MAP for all cases. Therefore, PSO has a better ability to find the optimum solution than the other two methods.

To demonstrate the excellent performance of implementing the PSO to estimate MAP, the three parts of Table 1 are represented in Figure 2, Figure 3, and Figure 4, respectively.

From Figures (2, 3, and 4), the implementation of PSO in Method 2 achieves more reliable results of MAP than Method 1 and Method 3.

FUTURE TRENDS

The suitable values of stroke volume, total peripheral resistance, and the coefficient of SV changing (A) will be investigated to achieve high accuracy, by testing more cases, and by using the multiobjective PSO technique.

Table 1. Means of MAP (mmHg), AD and APD for all cases

	Means of MAP (mmHg)				AD			APD		
No. of Cases	Real	Meth.1	Meth.2	Meth.3	Meth.1	Meth.2	Meth.3	Meth.1	Meth.2	Meth.3
1	63.45	63.32	77.16	108.16	0.13	13.71	44.71	0.21%	21.61%	70.47%
2	80.03	65.00	82.88	122.91	15.03	2.85	42.88	18.78%	3.56%	53.58%
3	75.34	58.62	67.98	88.94	16.73	7.37	13.59	22.20%	9.78%	18.04%
4	80.81	61.40	72.97	98.89	19.41	7.84	18.08	24.02%	9.70%	22.37%
5	72.98	62.65	75.59	104.56	10.33	2.61	31.58	14.15%	3.58%	43.28%
6	75.62	64.79	81.78	119.85	10.84	6.16	44.23	14.33%	8.15%	58.48%
7	72.28	63.54	77.74	109.54	8.74	5.46	37.25	12.09%	7.55%	51.54%
8	79.45	60.38	71.04	94.91	19.07	8.41	15.47	24.00%	10.58%	19.47%
9	75.64	61.08	72.34	97.56	14.57	3.30	21.92	19.26%	4.37%	28.98%
10	77.60	64.47	80.54	116.51	13.13	2.94	38.91	16.92%	3.78%	50.14%
11	74.21	64.65	81.21	118.30	9.56	7.00	44.09	12.88%	9.44%	59.41%
12	76.80	63.02	76.43	106.48	13.79	0.37	29.67	17.95%	0.48%	38.64%
13	81.20	64.99	82.82	122.74	16.21	1.62	41.54	19.96%	1.99%	51.16%
14	75.32	64.08	79.26	113.24	11.23	3.94	37.92	14.91%	5.23%	50.35%
15	71.25	59.36	69.23	91.34	11.89	2.02	20.09	16.69%	2.83%	28.20%
16	71.58	62.01	74.21	101.53	9.57	2.63	29.95	13.36%	3.68%	41.84%
17	69.75	64.85	82.09	120.68	4.90	12.33	50.93	7.02%	17.68%	73.01%
18	69.46	65.16	84.17	126.72	4.29	14.71	57.26	6.18%	21.18%	82.45%
19	84.54	63.85	78.57	111.53	20.70	5.98	26.99	24.48%	7.07%	31.92%
20	72.54	64.23	79.72	114.41	8.31	7.18	41.87	11.45%	9.90%	57.73%
Average	74.99	63.07	77.39	109.44	11.92	5.92	34.45	15.54%	8.11%	46.55%

Figure 2. Means of MAP

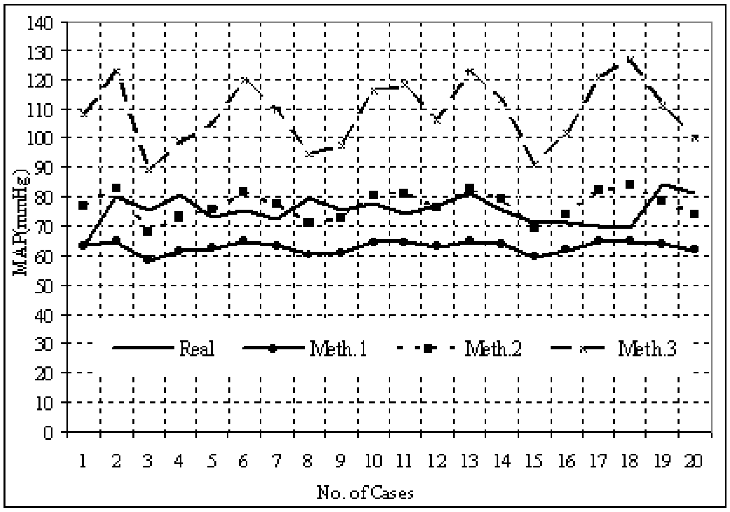


Figure 3. Absolute difference

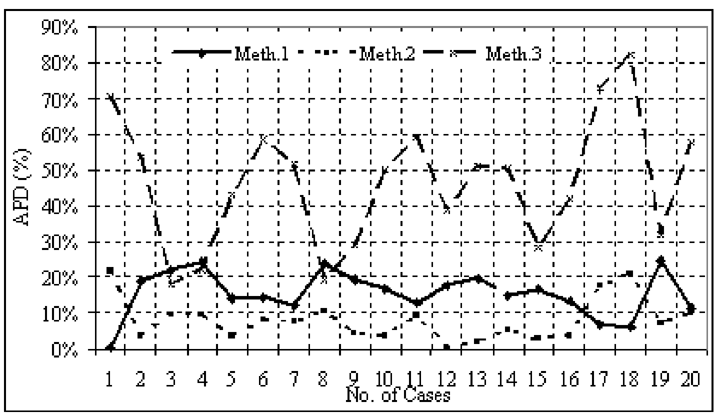
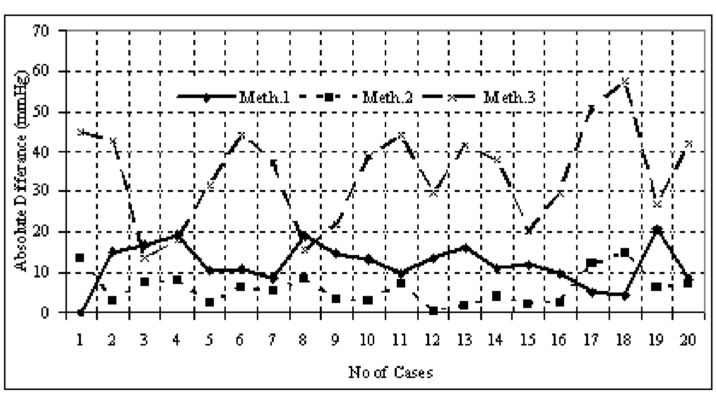


Figure 4. Absolute percentage difference



B

CONCLUSION

The relationship between HR and MAP is nonlinear (Shaltis et al., 2005). A lot of efforts were made to simplify this relation. In this article, a new algorithm was developed to estimate MAP, utilizing HR, stroke volume, total peripheral resistance, and considering SV changing influence; this consideration was investigated mathematically, and by the Particle Swarm Optimization technique. In addition, the suitable value of SV and TPR were investigated by PSO.

This new algorithm was implemented on 20 cases of mimic database (refer to Mimic Database Web site); 10 cases for training to investigate the suitable values of these factors, and all cases for verification and examining the performance of these techniques.

The estimated MAP by these mathematical models are compared with real MAP, and shown in tables and figures. The estimated MAP by Equation 3 with substituting A of PSO (Method 2) is achieved for best results.

In addition, Particle Swarm Optimization achieved an optimal value of stroke volume and total peripheral resistance, and the coefficient of SV changing percentage (A), which minimized the difference between real and estimated MAP.

The work of this article is part of our project to develop the Cardiovascular Parameters Long Term Monitoring System (CPLTMS), which depends on the PPG sensor.

REFERENCES

Bazett, H., & Dreyer, N. (1922). Measurement of pulse wave velocity. *Am. J. Physiol., 63*, 94–115.

Beecher. (2003). *Blood pressure monitoring*. Department of Medical Physics and Bioengineering, University College London. Retrieved February 2, 2008, from http://www.medphys.ucl.ac.uk/teaching/undergrad/projects/2003/group_03/history.html

Chen, W., Kobayashi, T., Ichikawa, S., Takeuchi, Y., & Togawa, T. (2000). Continuous estimating of systolic blood pressure using the pulse arrival time and intermittent calibration. *Med. Biol. Eng. Computer, 38*, 569–574.

Ganong, W. (2003). *Review of medical physiology, (internal ed.)* (pp. 621). New York: McGraw Hill.

Geddes, L., Hoff, H., & Badger, A. (1966). Introduction of the auscultatory method of measuring blood pressure--including a translation of Korotkoff's original paper. *Cardiovascular Research, 5,* 57–74.

Geddes, L., Voelz, M., Babbs, C., Bourland, J., & Tacker, W. (1981). Pulse transit time as an indicator of arterial blood pressure. *Psychophysiology, 18*, 71–74.

Kasirajan, V., Simmons, I., King, J., Shumaker, M., DeAnda, A., & Higgins, R. (2002). Technique to prevent limb ischemia during peripheral cannulation for extracorporeal membrane oxygenation. *Perfusion, 17*, 427–428.

Kennedy, J., Eberhart, R., & Shi, Y. (2001). *Swarm intelligence*. San Francisco: Morgan Kaufmann Publishers.

Klabunde, R. (2006). *Cardiovascular physiology concepts*. Retrieved February 2, 2008, from http://www.cvphysiology.com/Cardiac%20Function/CF002.htm

Lansdown, M. (1957). A method of using induced waves to study pressure propagation in human arteries. *Circulation Research, 5*, 594–601.

Ma, T., & Zhang, Y. (2005, September 1–4). A correlation study on the variabilities in pulse transit time, blood pressure, and heart rate recorded simultaneously from healthy subjects. In *Proceedings of Engineering in Medicine and Biology Society 27th Annual International Conference* (pp. 996–999). Shanghai, China.

Mackay, R. (1964). The application of physical transducer to intra-cavity measurement, with special application to tonometry. *Medical Electronic Biological Engineering, 2,* 3–17.

O'Brien, E., & O'Malley, K. (1981). *Essentials of blood pressure measurement with a foreword by George E. Burch Edinburgh*. New York: Churchill Livingstone.

Poon, C., & Zhang, Y. (2005, September 1–4). Cuffless and noninvasive measurements of arterial blood pressure by pulse transit tim. In *Proceedings of Engineering in Medicine and Biology 27th Annual International Conference* (pp. 5877–5880). Shanghai, China, 2005.

Piston, D., & Stradling, J. (1998). Value of beat-to-beat blood pressure changes, detected by pulse transit time, in the management of the obstructive sleep apnoea/ hypopnoea syndrome. *Europ. Respirat. J., 12*, 685–692.

Shaltis, P., Reisner, A., & Asada, H. (2005, September 1–4). Calibration of the photo plethysmogram to arterial blood pressure: Capabilities and limitations for continuous pressure monitoring. In *Proceedings of Engineering in Medicine and Biology Society 27th Annual International Conference* (pp. 3970–3973). Shanghai, China.

Yan, M., & Zhang, Y. (2005, September 1–4). Non-invasive estimation of blood pressure using photo-plethysmographic signals in the period domain. In *Proceedings of Engineering in Medicine and Biology Society 27th Annual International Conference* (pp. 3583–3584). Shanghai, China.

Zomaya, A. (2006). *Handbook of nature-inspired and innovative computing: Integrating classical models with emerging technology*. London, England: Springer Verlag.

KEY TERMS

Arterial Cannulation Method: The oldest method used to measure BP, by inserting a catheter on subject's artery.

Auscultatory Method: The most common method to measure BP by using sphygmomanometer.

Heart Rate: The rate of heart beating during a minute.

Mean Arterial Pressure: The mean of the force of blood flow against the artery walls.

Particle Swarm Optimization: A natural-based technique which relays on a certain insight concerning on persons actions and cognitions.

Photo-Plethysmography: The new method to measure the blood flow within the artery by PPG sensor.

Stroke Volume: The volume of the blood pumped by heart.

Total Peripheral Resistance: The summation of all body vessels' resistance against blood.

B

Bone Strength Assessment Based on CT Images Using a Network Spring Model for Radiation Therapy Patients

Yan Song
University of Houston, USA

Boris Mueller
Memorial Sloan-Kettering Cancer Center, USA

Chandra Burman
Memorial Sloan-Kettering Cancer Center, USA

Borys Mychalczak
Memorial Sloan-Kettering Cancer Center, USA

Yulin Song
Memorial Sloan-Kettering Cancer Center, USA

INTRODUCTION

In radiation therapy, especially using intensity modulated radiation therapy (IMRT), a high dose of ionizing radiation is delivered to the patient and, therefore, the bone health is a major concern for young patients. What is more, many tumors, like soft tissue sarcomas, are located in close proximity to bone. High energy radiation unavoidably exerts negative impacts not only on bone physiological properties, but also on its mechanical and structural properties. As a result, bone fracture risk for these patients is increased significantly. Thus, it is necessary to develop a quantitative noninvasive screening technique to identify these high risk patients.

BACKGROUND

Recently, IMRT has been increasingly used as an adjuvant treatment in conjunction with surgical resection for soft tissue sarcoma. Clinical studies have shown that wide excision followed by irradiation has resulted in excellent local control. Because most tumors extend to bones, particularly long bones, it is, therefore, necessary to perform excision of periosteum to obtain a clear margin. However, periosteal stripping destroys the nutrient vessels to bone and, thus, compromises the vasculature of the outer cortex. In addition, although IMRT is superior to conventional radiation techniques in critical organ sparing, for certain tumor sites such as soft tissue sarcoma, it still delivers a significantly high dose to bones. This affects bones negatively in many aspects. Hematopoietic cells of bone marrow are extremely sensitive to radiation. Clinical data have indicated that radiation dose at the level of ~ 1 Gy can produce bone marrow suppression. Lethal or sublethal damage to stem cells caused by radiation therapy prevents them from developing into differentiated myeloid and lymphocytic cells. Furthermore, osteoblasts and mesenchymal stem cells are also radiosensitive. Irradiation can reduce metabolic bone formative activity. It has been reported that as low as 8 Gy can inhibit the formation of heterotopic ossification.

The most common type of radiation treatment related to bone fracture is a stress fracture. The fracture risk rate may be as high as 24% in sarcoma patients who have undergone periosteal stripping and received chemotherapy. Thus, it is clinically relevant and significant to have a viable screening tool to identify those patients with high risk for IMRT treatment induced fracture. Since most sarcoma patients receive a follow-up CT scan upon completion of IMRT treatments to assess the treatment response, it would be very meaningful and cost-effective to develop a technique to assess bone fracture risk based on CT scans. Currently, bone densitometry is the most widely accepted method to detect bone fragility and to assess bone fracture risk. The established technique for measuring bone mineral density (BMD) is dual-energy X-ray absorptiometry (DXA). However, studies have shown that BMD is not the only determinant for bone fracture and other factors also contribute to bone fracture risk in other ways. *In vitro* experiments have revealed that bone strength depends not only on bone mass, but also on several mechanical and structural properties of bone, such as Young's modulus, bone microstructure, and geometric characteristics. These factors are not directly linked to BMD and, therefore, are not accounted for by dual-energy X-ray absorptiometry. In this article, we present a mathematical model for quantitatively assessing bone strength. The model, termed the spring-network model, takes these mechanical and structural factors into account and could provide a potential screening tool for predicting IMRT treatment related bone fracture risk.

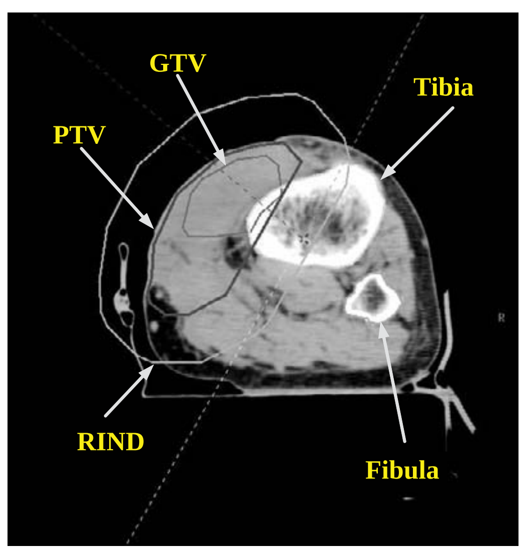
METHODOLOGY

IMRT

Since 2000, IMRT has emerged as the most important radiation therapy modality for soft tissue sarcoma treatment. Like conventional techniques, IMRT also employs the use of ionizing radiation to control and destroy malignant cells. The process of IMRT treatment planning consists of patient simulation, CT scan, target volume delineation, plan optimization, multileaf collimator (MLC) leaf sequence generation, and final dose calculation (Song, 2004; Xing, Wu, Yang, Boyer, 2004). The patient is first simulated on a simulator to determine the iso-center position and the possible treatment beam angles. In theory, the iso-center should be placed at the geometric center of the target. The goals are to fully use the central portion of the beam that posses most desirable beam properties and to minimize the probability of violating the MLC hardware constraints. Once the simulation is completed, the patient is CT scanned in the same position as in the simulation in a customized thermoplastic mold to minimize patient movement during the imaging process. At our cancer center, CT images of 2.5 mm slice thickness are acquired over the entire treatment region for treatment planning purpose. The planning target volume (PTV) is delineated on the CT images by the radiation oncologist. The PTV is created by adding a certain amount of margin around the gross tumor volume (GTV) to account for microscopic extension of the disease and treatment setup uncertainty. To achieve a better target dose coverage, a new structure named PTV_X is created by extending PTV superiorly and inferiorly one slice. In addition, a tuning structure, RIND, is also generated by expanding PTV_X by 2 cm axially to fine tune the dose distribution and to remove radiation hot spots from normal tissues. The treatment planner delineates the involved bones because they are the organ at risk in this setting and need to be spared.

Based on our clinical experience, a four or five-beam plan is necessary to create a concave dose distribution around the long bones. In selecting beam angles, all the beams enter the target from one side of the leg to spare a longitudinal strip of soft tissue. In addition, efforts are made to ensure that no beams enter or exit through the uninvolved leg. Furthermore, all beams should not pass through or be close to genital organs for patients with a tumor in the upper thigh and pelvic region. The collimator of each beam should be rotated to conform to the contour of the PTV for the best beam shaping. Figure 1 shows the volume definitions for a representative soft tissue sarcoma IMRT plan. The IMRT plan is computed using an in-house treatment planning system. Given a set of dose constraints, each IMRT plan is optimized based on a quadratic objective function. Once optimal beam intensity profiles or maps are obtained, MLC leaf sequences are generated using the

Figure 1. Structure definitions used in IMRT inverse treatment planning optimization. In this particular case, the tumor was in the right leg



dynamic MLC (DMLC) technique. Based on these leaf sequences, the final dose distribution is then computed using a pencil beam algorithm. If the criteria for plan acceptance are not met, a trade-off between the target dose coverage and normal tissue constraints would have to be made. The patient shown was treated with six MV photons on a Varian Clinac 21EX (Palo Alto, CA) equipped with a 120-leaf MLC. Figure 2 shows the MLC aperture projected on the digitally reconstructed radiograph (DRR) at a beam angle of 210^0.

Spring-Network Model

The quantitative assessment of bone strength is based on a novel mathematical model called spring-network model (Song, 2004), where the bone specimen is constructed by many elastic elements (or called occupied cubic voxels in the 3-dimensional CT image space with a grid of 512x512xslice number). Every voxel consists of 24 springs (Ladd & Kinney, 1997; Ladd, Kinney, & Breunig, 1997). Twelve of them have an elastic constant k along 12 sides of the cubic voxel, the remaining 12 have an elastic constant 2k along face-diagonals (Ladd et al., 1997), as shown in Figure 3. Under this construction scenaior, the spring network is isotropic and stable in configuration, although no cubic-like network is precisely isotropic mathematically (Ladd & Kinney, 1997).

The following boundary conditions are applied to this model: when the network is subjected to a vertical compression ς, all nodes on the top surface are free to move in horizontal direction uniformly in their displacement; vertical motion is not allowed. All nodes on the bottom surface move within horizontal plan and no

Figure 2. A representative MLC aperture at a beam angle of 210^0. The open area was part of the leg that needed to be treated. The other areas were blocked by the MLC.

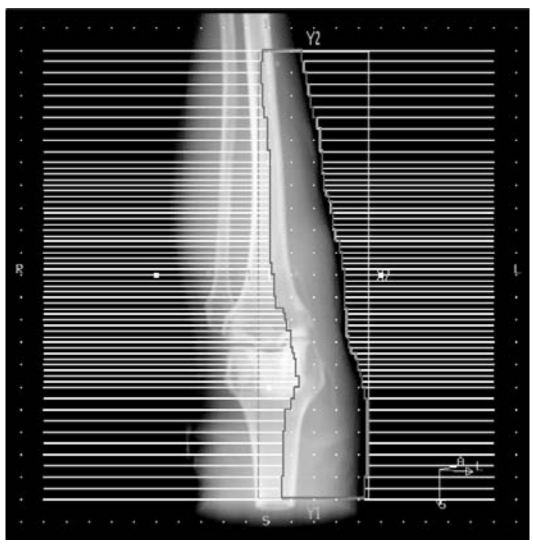
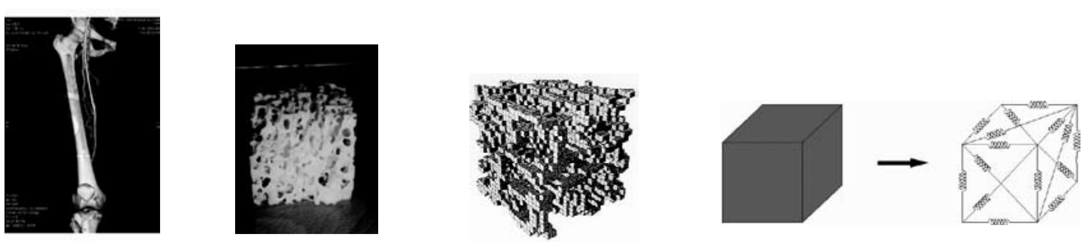


Figure 3. A bone specimen is represented by microscopic cubic building blocks through a series of modeling, where each block (or voxel) is described as a 24 linear spring system positioned in the CT image Cartesian coordinator



vertical motion is permitted. All other nodes in the bulk can shift in the way determined by all the elastic forces exerted by all its neighbors through linear mechanical law. The dynamical behavior of the network under external compression ς is described by classical particle in each node with mass *m* and viscosity constant κ. The mass *m* is one eighth of the total mass of each voxel if this node is at one of the vertices with 8 neighbors. If this node is at the edge or boundary of the lattice network, the number of occupied voxels in neighbor is less than 8, the mass *m* is then scaled by a factor $s < 8$ from voxel mass $sm/8$. In the same consideration, the elastic modulus kg_{ij} connecting nodes *i* and *j* should be scaled by a factor, depending on whether its neighbor is 8 or less. The Hamiltonian of the network is

$$H = \sum \frac{p_i^2}{2m_i} + \frac{k}{4} \sum g_{ij} [(u_i - u_j).\hat{x}_{ij}]^2$$

where p_i denotes the momentum of the mass at node *I*, u_i its displacement, and $\hat{x}_{ij}$ the unit vector in the direction from node *i* to node *j*. The equilibrium configuration of the nodes in the network subject to a compression ς is determined by minimizing potential energy *H* given above. This minimization is realized by the conjugate-gradient method under the certain choice of initial conditions and model parameters, otherwise, the equilibrium configuration can not be reached within finite steps of calculation.

The network fracture is simulated by the failure of individual elastic elements when a threshold strain η is reached: when the strain of a voxel is bigger than η, this voxel will be set to 0 for its mass and elastic modulus and this voxel is no longer occupied. We set a constant fracture-strain threshold for every voxel in this model because experiments indicated that for a given anatomical location of human trabecular specimen, the fracture-strain is nearly independent of human age (Gao, Jager, Artz, & Fratzl, 2003; Liebachner, 2005; Rajapakee et al., 2004), thus independent on the level of mineralization. The strain is calculated from the largest eigenvalue of the strain tensor (all direction components are included in this tensor), thus, in this model, the failure strain is independent of the degradation of the network model. It is shown that if the network contains large porosity, the fracture strain will reduce because of the large likelihood of fracture size (Gao et al., 2003; Rajapakee et al., 2004). If there is enough amount of new stress pathway, the network exhibits certain plasticity no matter how individual voxels are assumed to be brittle in the model assumption.

Bone Strength Assessment

To assess the bone strength, CT images with bone cross section are converted from DICOM format to TXT format to obtain the voxel information, such as voxel size and CT number. A three dimensional Cartesian coordinator is constructed with a resolution of 512 x 512 x *h*, where *h* is the number of CT slices. If a voxel value is not zero, its position is identified as (*x*, *y*, *z*) and it is replaced by an elastic element (24 linear spring boxes). Since a voxel value is directly related to its CT number, thus, it is proportional to the mass at that point of bone specimen. We assign a voxel value as its mass and elastic modulus to the corresponding elastic element in the assessment.

As a given model parameter, all nodes on the top surface vertically shift a constant value. It represents a mechanical load applied to the network and all nodes on the bottom layer are allowed to move in a horizontal plane. The vertical motion is not allowed. The displacements of all the other nodes in the bulk are adjusted accordingly and finally a stable configuration is reached with minimized Hamiltonian. If the network is not uniform, for example, if some porosity exists, or some voxels have a small elastic modulus, the local strain will be larger than the threshold that we preselected in model parameter. This voxel will be eliminated. Consequently, the neighbor voxel will adjust its strain until a stable configuration is reached again. In the stable configuration, a response force proportional to the displacement of position on top layer nodes is derived according to Hooke's Law. If vertical compression h increases step by step, a series of response forces are derived and a strain-stress curve is obtained. In this process, if more and more voxels are eliminated, eventually an avalanche occurs and the response force will drop rapidly until the network completely elapses. Figure 4 shows two representative CT images used in the calculations. Figure 5 shows the four 3-D bone specimens reconstructed from a series of 2-D CT images. Figure 6 shows the response stress under the external loading for four soft tissue sarcoma cases prior to IMRT treatment. The slopes indicate the stiffness (Young's modulus) of the bones.

Figure 4. Two representative CT images showing femoral bones

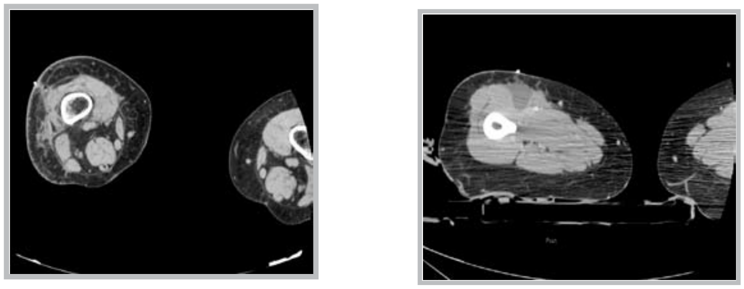


Figure 5. Bone specimen of the four representative cases used in the study. The bone specimens were reconstructed from the 2-D CT images as shown in Figure 4.

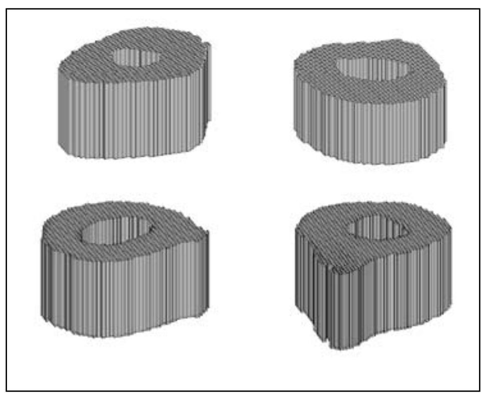
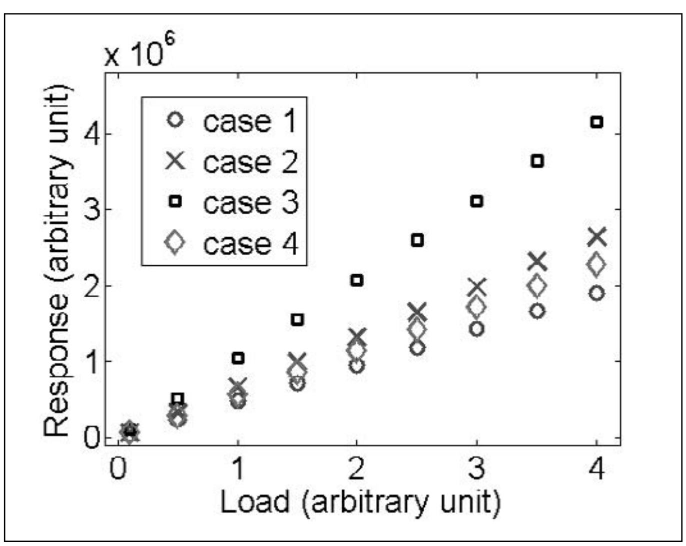


Figure 6. Mechanical response as strain applied at the top layer of the network. The slop indicates the stiffness of the bone specimen.



FUTURE TRENDS

Despite significant technological advances in computed tomography, anatomic definition of bone at a microscopic level remains limited. Currently, commercial diagnostic CT scanners have a nominal 512x512 in-plane resolution and a millimeter slice thickness. With a smaller field of view (FOV), a submillimeter in-plane resolution is achievable. However, it is still insufficient for quantitatively imaging the microstructure of bone at the micrometer level. As the detector size becomes smaller, the detecting efficiency improves, and faster reconstruction algorithms are developed, commercial ultra-high resolution CT scanners will eventually become available, leading to CT Microscopy. Indeed, high-resolution MicroCT has been developed and used in ultrahigh resolution imaging in the research laboratory on small animals and pathologic specimens. With CT Microscopy, the spring-network system will be able to model the mechanical properties and geometric structure of bone more precisely, thus providing a more accurate quantitative estimate of bone fracture risk. In addition, *in-vitro* experiments will be performed to calibrate our model system so that an absolute bone fracture factor can be calculated.

CONCLUSION

Based on the results of our preliminary study, we find that the spring-network model is an easy and practical approach to assessing bone strength. As shown in Figure 6, the mechanical response (stress) of the bone sample is linearly proportional to the applied load (with increment of one unit displacement at the top layer). The units in Figure 6 are arbitrary, so only the change in the curve direction or the change in the slope between the pre- and post-radiation treatments is important. By determining the slope difference between the pre- and post-radiation treatment loading-response curves, we will be able to tell the effects of IMRT treatment on the bone strength and its fracture risk. We believe that this model is a useful screening tool for bone strength assessment and has potential applications in radiation oncology as an increasing number of cancer patients receives IMRT treatments.

REFERENCES

Gao, H., Ji, B., Jäger, I.L., Artz, E., & Fratzl, P. (2003). Materials become insensitive to flaws at nanoscale: Lessons from nature. *Proc. Natl. Acad. Sci., 100,* 5597-5600.

Ladd, A.J.C., & Kinney, J.H. (1997). Elastic constant of cellular structures. *Physica A., 240,* 349-360.

Ladd, A.J.C., Kinney, J.H., & Breunig, T.M. (1997). Deformation and failure of cellular material. *Phys. Rev. E.***,** *55,* 3271.

Liebschner, M.A.K., Müller, R., Rajapakee, C.S., Wimalawansa, S.J., & Gunaratne, G.H. (2005). Testing two-predictions for fracture load using computer models of trabecular bone. *Biophys. J., 89,* 759-767.

Rajapakse, C.S., Thomsen, J.S., Espinoza-Ortiz, J.S., Wimalawansa, S.J., Ebbesen, E.N., Li, M., et al. (2004). An expression relating breaking stress and density of trabecular bone. *J. Biomech., 37,* 1241-1249.

Song, Y., Boyer, A., Pawlicki, T., et al. (2004). Modulated electron radiation therapy—Emerging technology. In *Intensity modulated radiation therapy—A clinical perspective*. Lewiston, NY: BC Decker.

Song, Y., Liebschner, M.A., & Gunaratne, G.H. (2004). A study of age-related architectural changes that are most mamaging to bones. *Biophys. J., 87,* 3642.

Xing, L., Wu, Q., Yang, Y., & Boyer, A. (2004). Physics of IMRT. In *Intensity modulated radiation therapy—A clinical perspective.* Lewiston, NY: BC Decker.

KEY TERMS

Intensity Modulated Radiation Therapy (IMRT): A new radiation therapy technique based on the concept of inverse treatment planning. By optimizing a set of dose-volume constraints and dose limits, the radiation intensity can be finely modulated so that the dose distribution matches the geometrical distribution of the target volume. Therefore, the maximum dose of radiation can be delivered and the side effect can be minimized.

Strain: The relative deformation (size or shape) of a material under stress.

Stress: The internal distribution of force over a unit area. It is the interaction response to the external loading.

Voxel: The smallest volume element in a 3-dimensional image space determined by the image resolution. It is the basic image element of the object reconstructed from a series of 2-D CT images.

Young's Modulus: The ratio of material stress to strain (when strain is not very large). It is a measure of material stiffness.

The Canadian Health Record Interoperability Infrastructure

Jens H. Weber-Jahnke
University of Victoria, Canada

INTRODUCTION

Countries around the globe are struggling with the rising cost of delivering health care. In the developed world, this trend is enforced by aging demographics and emerging forms of expensive medical interventions. Disease prevention, early disease detection, and evidence-based disease management are key for keeping health care systems sustainable. Electronic information management has been recognized as a central enabler for increasing the quality of health care while controlling the cost of delivering it. Secondary care facilities (e.g., hospitals) and laboratories have made use of electronic information systems for decades. However, the primary care sector has only recently begun to adopt such systems on a broader scale. The benefit provided by each system in isolation is limited since citizens generally receive their care from a multitude of providers. Health care information systems need to interoperate in order to enable integrated health information management and consequently attain the declared qualitative and economic objectives. Many industrial countries have begun to create common infrastructures for such an integrated electronic health record (EHR) (Blobel, 2006). Different approaches exist, ranging from centralized databases to highly distributed collections of mediated provider-based systems. This chapter describes the architecture of the Canadian infrastructure for health information management, which can be seen as a compromise between a fully centralized and a fully distributed solution. While in Canada the delivery of health care is a matter of provincial territorial authority, the health ministers of all provinces and the federation have created a joint organization called Health Canada Infoway with the mandate to develop an architecture for and foster implementation of a joint interoperability infrastructure for EHRs in Canada. The second major version of this architecture has now been released, and provinces have begun to implement it. The solution is based on the paradigm of a service-oriented architecture (SOA) (Erl, 2004) and embraces a range of domain-specific and technical standards. It leverages and integrates existing investments in health information systems by making them available through interface standards-conform interface adapters. The Canadian EHR architecture has received attention beyond the Canadian context. This chapter reports on this architecture, its enabling technology paradigms, experiences with its implementation, and its limitations.

THE CANADIAN HEALTH CARE SYSTEM

Canada is a federation of 10 provinces and three territories. Funding and administration of health care services falls under the jurisdiction of these individual provinces and territories. Federal legislation (the Canada Health Act) provides a common set of conditions under which provinces and territories receive funding for health services. These conditions include *universal access* to health care for all Canadians independently of their income. Moreover, they require portability of insurance coverage among provinces such that Canadians remain entitled to medical care while moving among provinces. The Canada Health Act covers basic health services only, including primary care and hospital services. Many other services such as dental care, optometry, and prescription medications are excluded. Private insurance plans are available for purchase to cover these services. They are also a typical component of employee benefit packages.

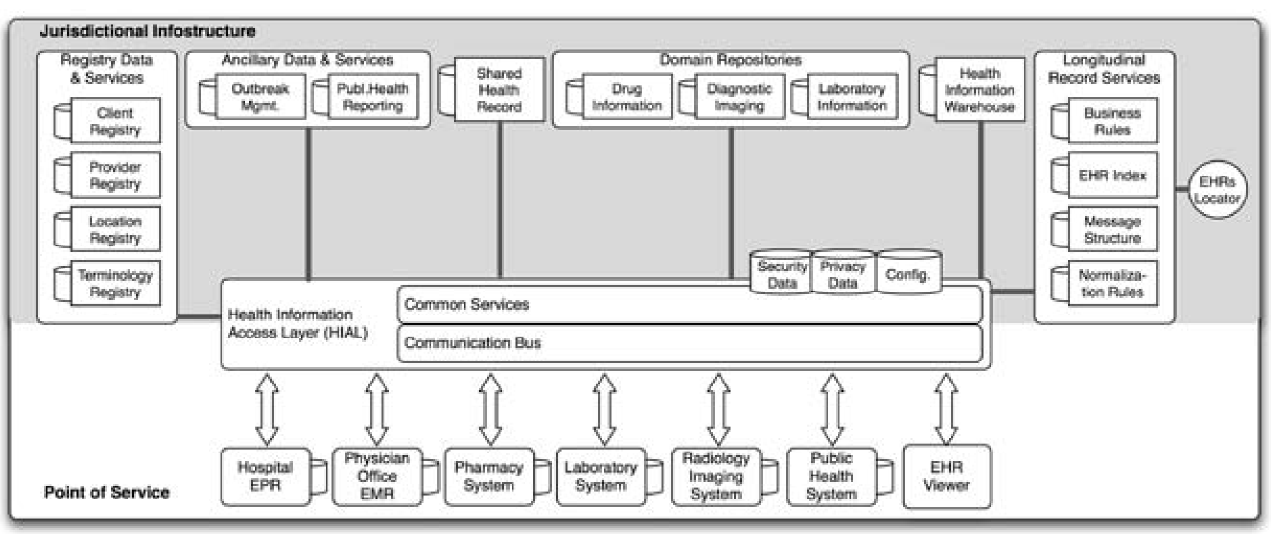
There are approximately 60,000 physicians practicing in Canada, half of them in primary care, the other half as specialists. The Canadian health care system has been slow to adopt new information and communications technologies (ICTs) (Newbery, Gelhorn, Gutkin, Renaud & Challis, 2000). This is particularly the case for the primary care sector. While ICTs are increasingly being used for administrative services such as billing, scheduling, and managing basic demographic data, the majority of medical records is still kept in paper files. At the turn of the millennium, Canada launched initiatives focused on removing the barriers of making the transition to a more efficient and modern health care system. An example is the *Primary Care Transition Fund*, which has invested $800 million over six years (2000–2006) for renewing the health care systems across the country. The Canadian Health Accord released in 2004 recognizes that "electronic health records and telehealth are key to health system renewal, particularly for Canadians who live in rural and remote areas."

The decentralized nature of the Canadian health care system has contributed to the development of a large number of heterogeneous medical information systems. A recent estimate puts this number at 40,000 (Allas, 2006). Integrating all these systems to achieve more efficient and effective use of medical information poses a significant challenge. In order to facilitate this integration, the health ministers of all provinces, territories, and the federal government have jointly formed an organization called *Health Canada Infoway* (Infoway, 2006) with a mandate to facilitate the development and adoption of interoperable electronic health (EHR) record solutions. Founded in 2001, Infoway has the goal of connecting at least 50% of all Canadians to an interoperable EHR by the end of 2009.

ARCHITECTURE

The Canadian strategy for achieving an interoperable EHR pivots on a joint information system reference model called the EHR solution blueprint architecture (EHRs blueprint). Infoway published the first version of the EHRs blueprint in 2003. Its second and final major version update was released in 2006 after extensive consultation with various stakeholder groups. This major revision adds concepts for *Telehealth* and *Public Health Surveillance* as well as extensive requirements and design specifications for security and privacy concerns. According to Richard Alvarez, Infoway's president and CEO, the EHRs blueprint provides the Canadian framework for accessing and consolidating health information and presenting it in a way that meets the needs of health care professionals such as GPs, specialists, nurses, and pharmacists across various care settings, including hospitals, emergency rooms, clinics, home care, and geographical distances (Infoway, 2006). The EHRs blueprint is based on the *Service-Oriented Architecture* (SOA) paradigm, which promotes loosely coupled and standards-based component interfaces (Erl, 2004). The following diagram gives a conceptual overview of the EHRs blueprint architecture, which consists of the *jurisdictional infrastructure* part (also called the *EHR Infostructure – EHRi*) and the actual application components used at the *point of service* (PoS).

Figure 1.



The EHRi consists of the following:

- **Registry services** to manage and provide directory information about actors, resources, and terminology in the EHRs, including a registry of health care providers, patients, and locations, as well as standard terminologies to describe diseases, acts, and other clinical information.
- **Domain repositories** to store and provide clinical information about patients pertinent to a particular domain (e.g., diagnostic imaging, laboratory data, drugs).
- A **shared health record** to manage and provide information about patient health histories.
- **Longitudinal record services** to orchestrate and track distributed care processes, including access, updates, and location of health data.
- A **health information data warehouse** for collecting health information for the purpose of public health surveillance.
- A **health information access layer** (HIAL) that forms the middleware between the EHR infostructure and individual point of service (PoS) systems, including a communication bus and common services (e.g., for privacy and security).
- A generic class of **ancillary services** that rely on, need, or complement EHR data (e.g., communicable disease reporting, outbreak management, and wait-list monitoring).

The EHRs blueprint architecture can be considered a compromise between a fully decentralized federation of autonomous medical record systems and a fully centralized information system concept. It recognizes that the individual PoS information systems will keep most of their autonomy hosting patient-centric health information. However, the Shared Health Record (SHR) component of the EHRi assumes the role of the *definitive* medical summary that is shared among medical service providers. Moreover, the SHR forms a longitudinal record that should cover a patient's health history "from womb to tomb."

Geographically distributed EHRs implementations subscribe to a common *EHRs Locator* service component, which enables them to locate other EHRs implementations that maintain health data about a particular patient. For example, a patient who has accessed health services provided by four health authorities (running their own EHRs) will have his or her clinical information residing in four EHRs implementations. However, any such storage fragmentation of personal health information is hidden from the point of view of the individual PoS applications accessing any EHRs. The Longitudinal Record Services (LRS) of each EHRs use the EHRs Locator to automatically locate all relevant storage locations in order to assemble a cohesive health record.

While the HIAL facilitates integration of existing PoS applications, their integration still requires adaptation effort. Therefore, providers can also use a dedicated *EHR Viewer* to access the EHR data while their PoS application is not yet fully integrated with the HIAL or when their practice does not yet use an electronic medical record system.

INTERACTION MODEL

Various PoS systems communicate updates to patient health records by means of the SHR. The EHRi does not support direct peer-to-peer communication between different PoS applications. The EHRi plays a passive role when it comes to interacting with PoS applications. This means that it does not actively *push* any information to or *pull* information from connected PoS applications. Rather, PoS applications use the EHRi to retrieve health data or push local updates to the shared infrastructure. When a patient accesses the health care system at a particular PoS, the information system installed at that location queries the SHR for the most up-to-date health record for this patient. If the patient has never accessed the health care system at this particular PoS, a new local health record is created based on the record provided by the SHR. Otherwise, the local record is updated by any SHR updates since the patient's last visit at this location, or the local data and the SHR data are presented side by side to the clinician. Clinical data captured at the PoS application may be promoted to the EHRi level in order to update the SHR as a result of the new patient encounter.

INTEROPERABILITY STANDARDS

The interoperability standards used in the EHRs blueprint are based on the Health Level Seven (HL7) V.3 messaging architecture (Beeler, 1998). Some messages use coded data, while others use unstructured information (i.e., free text). The standard uses coded messages for updating registries (e.g., service delivery locations) as well as messages for recording clinical observations and (potentially composite) professional services. Semistructured (partially uncoded) messages are used to record referrals, clinical documents (observations), discharge and care summaries, and general health conditions, and to retrieve a patient's SHR summary. Coding systems used are SNOMED-CT, ICD-10 (with Canadian enhancements), and the Canadian Classification of Health Interventions (CCI) (Chute, 2000).

SECURITY AND PRIVACY

Infoway released a detailed specification of privacy and security requirements in 2005 that considered and complied to national, provincial, and territorial privacy legislations, policies, and standards. These requirements are partially of a technical and partially of an administrative nature. They are also categorized regarding their scope (i.e., whether they pertain to Point of Service (PoS) applications connecting to an EHRi, or whether they pertain to organizations hosting EHRi components). While Canada has federal privacy legislation (*Personal Information Protection and Electronic Documents Act*), most provinces have their own legal frameworks. Therefore, EHRs requirements are defined conditionally. For example, certain aspects of dealing with consent are not consistently defined in the requirements specification. As such, it is not defined whether participation in the EHRi is mandatory or whether there are choices to opt in or opt out. In 2005, Infoway further published a conceptual architecture of privacy and security services that are considered necessary and sufficient to enable implementation of the specified requirements. These services include anonymizing filters, encryption services, services to ensure integrity and perform authorization, authentication, and auditing. One important principle of the privacy and security architecture is that it uses separate repositories for storing medical information and personal identifiable information. Both data sets are linked up only by a reference identifier that has no other purpose or semantics.

IMPLEMENTATION EXPERIENCES

Different jurisdictions may implement different technical realizations of the aforementioned EHRs architecture according to their needs. The Infoway blueprint defines the EHRs architecture only on the conceptual level and logical levels, but leaves the definition of the physical level up to the individual jurisdiction. It is expected that as long as different EHRs implementations use the same standards, they will be interoperable to ultimately create the envisioned pan-Canadian EHR. At the same time, Infoway has funded and actively worked with public and private partners to develop *reusable* components for this architecture with the objective to "*build once but use many times.*"

Several components of the EHRs blueprint architecture have been developed in various jurisdictions throughout Canada. Early implementation efforts have concentrated on developing registry services (e.g., patients, providers, locations, drugs, etc.) and domain repositories (e.g., digital imaging, laboratories). Infoway has started funding initiatives for developing the infrastructure components that integrate this information. Currently, Infoway is funding and actively working with more than 150 projects.

Several Canadian provinces have already developed infrastructures for particular aspects of the planned EHR (e.g., shared drug information systems). These projects have influenced and have been influenced by the two subsequent versions of Infoway's national EHRs blueprint. They provide a first limited evaluation of the EHR blueprint. To date, the arguably most difficult component, the component that provides the *Longitudinal Record Services* (LRS), still remains to be implemented in most jurisdictions. While the current second version of the EHRs blueprint is considered final when it comes to the publication of a cohesive document, Infoway has created a joint artefact repository for sharing experiences made with implementing the EHRs blueprint and for refining the specifications on an ongoing basis.

DISCUSSION AND COMPARISON

The Canadian EHR infrastructure is most similar to the Australian *HealthConnect architecture*, which is expected to be implemented by 2010 (Wooding, 2003). Both models advocate a federated model leveraging existing PoS applications and implement a distributed repository for shared health record data. They also use a similar set of standards. Patient participation in HealthConnect is voluntary on an opt-in basis.

In Europe, the *NHS Care Records Services* project in the United Kingdom and the *Dossier Médical Personnel* (DMP) in France implement a more centralized notion of the shared health record. The core database of the NHS project, called the *Spine*, went live in 2005, while the DMP is expected to go online in 2007 (Cross, 2006). These projects defer from Canada's EHR project in that they represent fully centralized repositories for all citizens rather than specifying a blueprint standard that is implemented separately across different jurisdictions. The different approaches reflect on the general differences in the countries' health care systems. There have been significant concerns regarding privacy and security of these centralized solutions. Both systems make a default assumption of citizen participation. Privacy advocates demand for the right to opt out.

Other European countries such as Germany are using a different approach to tackle the interoperability challenge. They provide citizens with "smart" health cards to store and share health data. Germany started a trial run involving 10,000 patients in December 2006. Over the course of the project, the card will gradually store an increasing amount of information, including prescriptions, medications, and allergies, up to the complete electronic medical record. Patient participation in this system is mandatory, but patients can decide how much information is kept on their cards (BMG, 2006).

Countries lacking a cohesive public health care system have seen developments of various disparate EHR initiatives. In the United States, the most notable system on the national level has been developed by the Department for Veteran Affairs. The *Veterans Health Information Systems and Technology Architecture* (VistA) is a client-server system maintaining care records of approximately 4 million veterans. Being developed for a single organization, VistA puts more emphasis on developing the PoS application user interface than the EHRs, which mainly focuses on infrastructure. VistA includes an extensive GUI client called the *Computerized Patient Record System* (CPRS) (Fletcher, Dayhoff, Wu, Graves, & Jones, 2001). Other implementations of EHRs exist for separate American health care organizations.

REFERENCES

Allas, A. (2006). Canada Health Infoway: EHRS Blueprint. Canada Health Infoway.

Beeler, G.W. (1998). HL7 version 3—An object-oriented methodology for collaborative standards development. *Int J Med Inform*, *48*(1-3), 151–161.

Blobel, B. (2006). Advanced and secure architectural EHR approaches. *Int J Med Inform*, *75*(3-4), 185–190.

BMG. (2006). *Die elektronische gesundheitskarte*. Berlin: Bundesministerium für Gesundheit, Referat Öffentlichkeitsarbeit. Retrieved from www.bmg.bund.de

Chute, C.G. (2000). Clinical classification and terminology. *Journal of the American Medical Informatics Association*, *7*, 298–303.

Cross, M. (2006). Information technology: Will connecting for health deliver its promises? *British Medical Journal*, *7541*, 599.

Erl, T. (2004). *Service-oriented architecture: A field guide to integrating XML and Web services*. Upper Saddle River, NJ: Prentice Hall.

Fletcher, R.D., Dayhoff, R.E., Wu, C.M., Graves, A., & Jones, R.E. (2001). Computerized medical records in the Department of Veterans Affairs. *Cancer*, *91*(s 8), 1603–1606.

Infoway. (2006). *EHRS blueprint—An interoperable EHR framework*. Montreal: Canada Health Infoway.

Newbery, P., Gelhorn, D., Gutkin, C., Renaud, C., & Challis, L. (Eds.). (2000). *Primary care and family medicine in Canada: A prescription for renewal*. Mississauga, ON: The College of Family Medicine in Canada.

Wooding, R. (Ed.). (2003). *Health connect draft system architecture*. Canberra: The HealthConnect Program Office, Dept. of Health and Aging. Australia.

KEY TERMS

Domain Repository: A component of an EHR Infrastructure that stores, manages, and persists a clinical subset of data pertaining to a specific domain.

EHR Infrastructure: A collection of common and reusable components in support of a diverse set of health information management applications. It consists of software solutions, data definitions, and messaging standards for the EHR.

Electronic Health Record (EHR): Provides a secure and private lifetime computer-based record of a patient's key health history and care within a health system.

Health Information Access Layer: A gateway that acts as an abstraction layer to separate PoS applications from the EHR Infrastructure. It is made up of service components, service roles, information models, and messaging standards required for the exchange of EHR data and the execution of interoperability profiles among EHR Services.

Health Information Data Warehouse: An information system to compile, aggregate, and consolidate EHR data for reporting and statistical or research analysis.

Point of Service (PoS) Application: PoS applications are used by authorized caregivers to view and navigate the EHR of a patient/client.

Registry: A directorylike system that focuses solely on managing data pertaining to one conceptual entity.

Challenges with Adoption of Electronic Medical Record Systems

Abirami Radhakrishnan
Morgan State University, USA

Dessa David
Morgan State University, USA

Jigish Zaveri
Morgan State University, USA

INTRODUCTION

Among health care information systems, past research has credited Electronic Medical Records (EMR) systems with offering the greatest potential for improving quality within health care environments. Benefits range from reducing errors to cutting overall health care costs. For instance, the utility of an EMR system will allow physicians' enterprise wide access to a patient's entire medical chart, monitor patients' care for possible drug interaction, proactively prompt doctor(s) with recommended treatment, provide clinical decision support, simplify record keeping, e-prescription, documented referrals, and reminders to patients and health care providers.

Despite these benefits and a defined movement to integrate EMR systems in medical outfits, adoption of EMR systems by health care professionals has been very slow (Audet, Doty, Peugh, Shamasdin, Zapert, & Schoenbaum, 2004; Burt, Hing, & Woodwell, 2005; Miller & Sim, 2004; Simon & Simon, 2006). According to the National Ambulatory Medical Care Survey Report (2005) only 25 % of office-based physicians are recorded as partial or fully using EMR systems. Nevertheless, interest to adopt EMR systems continues to be significant (Miller & Sim, 2004).

What accounts for the slow adoption of EMR systems? To answer, we must identify and address challenges associated with this process. A review of the recent practitioners, academic health informatics literature, and provisions of HIPPA Act of 1996 (Adler & Edsall, 2005; Audet et al., 2004; Baharozian, 2005; Edsall & Adler, 2005; Hough, Chen, & Lin, 2005; Lenhart, Loomis, Criswell, & Meggs, 2000; Miller & Sim, 2004; Retchin, Wenzel, &, 1999; Swartz, 2005; Valdes, Kibbe, Tolleson, Kunik, & Petersen, 2004) cite several barriers faced with the adoption process. Further analysis also suggests that the promises of successful EMR deployment will not be fully realized unless concerns linked to the EMR implementation process are alleviated. We investigated EMR adoption by conducting open ended interviews with EMR managers, vendors, and physicians to explore their experiences with their EMR implementation.

In this article, we present the results from our study. The next section highlights challenges associated with EMR adoption and use. We conclude by suggesting solutions geared towards lessening these challenges thereby clearing the path for successful EMR adoption and use.

CHALLENGES

Our meta-analysis identified several barriers experienced by professionals regarding EMR adoption and use. These challenges include: cost, difficulty in calculating return on investment, lack of education, physicians' and staff concerns, technology related concerns, inadequate complementary changes to organizational processes, lack of IT support, and lack of incentives.

Costs

EMR systems are costly. Many health care institutions cite cost as a primary prohibitive factor with adoption of EMR. There are high up-front installation costs and recurring expenses for operation and maintenance. During our interviews, the interviewees clarified that up-front costs range from $15,000-$60,000 per physician.

Initial set-up costs include purchase cost of hardware, software, network infrastructure, trainin, and workflow reorganization. Operation and maintenance costs include data conversion, ongoing training, hardware and software, and specialized IT support staff.

With such exorbitant costs and uncertainty regarding return on investments, we can assume that small companies may not find adoption of EMR systems feasible thus prohibiting implementation (Audet et al., 2004; Miller & Sim, 2004; Retchin et al., 1999; Winn, 2002).

Additional expenses are incurred during the EMR transition period due to physicians attending to fewer patients translating to decreased revenue.

Difficulty in Calculating Return on Investment

One of the major concerns with new projects for upper level management is financial payoff. Is this worth the investment? As mentioned earlier, the level of initial investment is high. There is an uncertainty over the size of financial benefits that may accrue over time (Audet et al., 2004; Miller & Sim, 2004).

Benefits obtained from EMR can be complex to measure with long pay back period. Most health care institutions lack the financial and operational analysis tools for an "uninformed" EMR buyer to make a competent decision on behalf of their organization.

Lack of Education

Most health care institutions lack knowledgeable personnel capable of evaluating and managing implementation EMR system for their organizations. EMR systems are complex with several modules and requiring special expertise.

During the implementation phase there is a need for a champion. The literature states projects without a champion are most likely to fail. A champion is one that promotes the benefits of EMR within the organization. He or she is a person capable of aligning the systems' functions with the needs of the health care organization. A champion can properly assist in defining the scope, allocating resources and preparing the organization for the transition that will come with EMR implementation.

Concerns of Physicians and Staff

The concerns of physicians and staff to use and manage EMR systems remain another major challenge. Physicians are very reluctant to adopt and use EMR systems. Physicians view their role primarily about patients, with automation secondary. Many fear the use of EMR systems will take them away from their primary duties.

Physicians and nurses also fear that managers could measure, compare, and evaluate the amount of time each professional spends on each task. Physicians and other medical staff are apprehensive that they may be reprimanded for "slacking off" or deviating from predetermined practice sequence. These concerns act as an impediment for adoption and use of EMR systems (Bar-Lev & Harrison, 2006). During our interviews, however, this factor did not emerge as one of the areas of concerns.

Physicians and staff are also reluctant to use EMR because they need to take time off their schedule for training on coding, documentation, and e-prescription capabilities of EMR systems (Berkowitz, 1997; Lenhart et al., 2000). Training requirements erode the initial enthusiasm for use among some users. During our interviews, a few physicians expressed that they spent a significant amount of time meeting the training requirements.

Physicians are also resentful with receiving clinical recommendations from EMR. They believe that after years of medical training, accepting recommendations from a computer information system is demeaning and a threat to their independent thinking.

Technology Related Concerns

There are several concerns related to EMR technology.

Technophobia

There are people within the health care clinics that are intimidated by technology. They simply panic when interacting with technology. Fears range from care providers perceiving that their jobs are at stake (EMR will replace or outperform them) to "what if I hit an incorrect key stoke and erase critical data from the system?" Such fears have crippling effects on EMR implementation.

Other Technology Concerns

Interoperability is another major concern. Interoperability refers to the ability of EMR system to facilitate exchange of patients' medical records across medical institutions such as hospitals, polyclinics, independent testing labs, and medical centers. This factor is important because there are more than 264 different software programs currently used in the U.S. and most of these programs are not compatible with each other (Retchin et al., 1999; Valdes et al., 2004). Poor interoperability results in poor electronic data exchange across institutions. Support staff and physicians get frustrated as they are forced to spend more time to manually enter data from external systems.

System difficulty is another technology related concern. This issue pertains to the difficulty associated with conversion of old paper based documents into electronic medical records (Berkowitz, 1997; Retchin et al., 1999). This factor need to be considered because many documents involve extensive handwritten content, some of which may have been generated by different health care professionals over the life span of the patient and some of the content is illegible for conversion. Moreover, the digital scanning process involved in conversion of these physical records to electronic documents is an expensive, time-consuming process, which must be done to exacting standards to ensure exact capture of the content.

Usability remains as a major concern. There are multiple screens, options, and navigational aids in the EMR system. Problems with EMR usability, especially documenting progress notes cause physicians to spend extra time learning effective ways to use the EMR (Audet et al., 2004; Miller & Sim, 2004). This issue is compounded by the fact that smaller medical institutions often lack financial resources to train physicians on usability.

Health care information is sensitive. There are concerns pertaining to data security and privacy (Baharozian, 2005; Rind & Safran, 1993). HIPPA Act (1996) mandates that all health care providers adhere to national standards while conducting electronic exchange of health data and emphasis should be laid on the security and privacy of health data. Even though it is recognized that EMR systems are better than paper based systems in terms of data security and privacy, these concerns still persist (Baharozian, 2005; Hsieh & Lin, 2006). There is a persistent fear that hackers can destroy sensitive patients' records.

Inadequate IT Support

With the implementation of any major information system as EMR, there is a need for IT support. Factors range from hardware, software, or network ongoing issues to training. EMR is also considered a reengineered process. Most professionals found in health care environments are medical professions with limited technology expertise. Many care providers express being more at ease using the system with IT support close by.

Inadequate Complementary Changes to Organizational Processes

EMR hardware and software cannot simply be used as "out of box" to reap superior business value. As Wade and Hulland (2004) point out, managers should effectively use outside-in and spanning resources such as IT management practices, IT change management, business systems thinking together with inside-out IT resources such as basic IT infrastructure and IT technical manpower. This simply means that information technology innovations and organizational process changes in tandem. Complementary changes needs to be done to the hospital processes such as patient registration, diagnosis, medical/surgical process, prescription generation and billing, and so forth. These complementary changes exact a great deal of time from physicians, nurses, front-office staff, and hospital managers. Physicians have to redesign their workflows (i.e., how they work in the exam room). Managers and support staff have to redesign office workflows. Practitioners' literature (Miller & Sim, 2004) point out that office-based physicians and solo physicians do not spend enough time in redesigning workflows. It is well known that hospitals that install sophisticated EMR systems before they have done the necessary organizational process changes are not going reap benefits from EMR investments. In our interviews, EMR managers agreed that extensive changes are needed in the management of work flow while trying to implement EMR systems.

Lack of Incentives

It is agreed upon that EMR systems are expensive and most medical professionals require extensive training to reap full benefits. The evidence from the literature suggests that given a choice many physicians will not use EMR systems. They view use of this system as time consuming and distracting from their primary duties. Without incentives by their institutions or government many health care professionals will not implement EMR systems.

SOLUTIONS

There are no simple anecdotes to overcome the challenges referenced above. We therefore propose a multifaceted solution. To expedite the process of EMR implementation, barriers experienced by health care professionals must be addressed at the individual, organization and governmental levels. On one hand we have risks feared by health care professionals and institutions. On the other hand we have a lack of incentives to entice businesses to adopt EMR.

We believe that interventions at the governmental and institutional levels can play a role in stimulating the adoption process. Each intervention increases the attractiveness of EMR use by hospitals and medical institutions.

Governmental Level Interventions

Certain interventions could be undertaken at the governmental level in order to spur the EMR adoption and use.

a. Federal and state governments could provide financial incentives, subsidies, and tax breaks to hospitals and medical institutions that adopt and use EMR systems. As we mentioned before, lack of financial resources is one of the major challenges faced by medical institutions and hospitals. This scheme of providing financial incentives would promote medical institutions and hospitals to invest and effectively use EMR systems. Audet et al. (2004) report that in European countries such as the UK and Sweden, physicians who invest in EMR systems receive government subsidies, which has spurred adoption of EMR systems. Such policies could be adopted by the U.S. federal government also. This scheme would also benefit the patients as EMR systems improve the overall efficiency of health care service providers.
b. As we mentioned before, system interoperability is one of the major technology related concerns. In order to address this concern, the U.S. Department of Health and Human Services could develop a process to harmonize EMR system applications and create industry-wide standards for operation and maintenance of EMR systems and exchange of health data. Vendors who sell EMR systems would be forced to come up with software applications that are compatible with each other. This would promote better standardization of the product, interoperability of EMR systems and information sharing of health records across medical institutions, hospitals and testing centers. This scheme would in turn promote usage of EMR systems. Miller and Sim (2004) report that the government has initiated some data exchange standards such as HL7, LOINC, and provisions of HIPPA Act of 1996 also mandate use of data exchange standards. We hope that more organizations use these data exchange standards so as to improve system interoperability.

Institutional Level Interventions

Hospitals and medical institutions could undertake the following interventions in order to promote adoption and effective usage of EMR systems by physicians and other staff.

a. **Provide performance incentives to physicians and other staff:** Hospitals and medical institutions could consider providing performance improvement based financial incentives to physicians and other staff. As mentioned before, many hospitals and clinics have reported that EMR systems have improved the throughput for diagnosis, treatment, and consultation, reduced the time taken for delivery of services, improved the speed and accuracy of medical insurance claims registration, reduced the number of medical coding and transcription errors, and improved the data integrity of patient's medical records and the institution's overall responsiveness to patient's needs. Hospitals could come up with a set of process based metrics (such as diagnosis throughput, number of medical errors, number of correct insurance claims registered, etc.) and evaluate the performance improvements associated with the use of EMR systems. Based on the extent of performance improvements achieved, physicians and other staff could be provided extra financial incentives. This would motivate the physicians and other staff to use EMR systems.

b. **Consider temporal aspect of technology payoff:** As IT business value studies (Brynjolfsson & Hitt, 2000; Devaraj & Kohli, 2000, 2003) suggest, top management of hospitals and medical institutions should consider the temporal aspect of technology payoff; that is, payoffs may not be realized instantaneously, but only after certain periods of time, which is usually 2-3 years. Hence it becomes important for top management to consider the lag effect when they conduct cost-benefit analysis of EMR systems.

c. **Top management support for use of advanced capabilities of EMR systems:** As mentioned before, the capability of EMR to document progress notes and facilitate audit trails create fear in the minds of physicians and nurses that managers could measure, compare and evaluate the amount of time each professional spent on each task. Physicians and other medical staff also fear that they might be reprimanded for "slacking off" or deviating from predetermined practice sequence. These fears act as an impediment for usage of EMR systems. Top management should lend its support to physicians and work to alleviate these fears. Top management should develop strong channels of communication and co-operatively work with physicians and other staff members. Also, physicians should be made to understand that coding, documentation, and audit trail capabilities should be viewed as positive features as they provide the required checks and balances in the health care delivery system.

d. **Increase thrust on training:** In order to work around technology usability issue, hospital management should place increased emphasis on training of system users (i.e., physicians and other office staff). Users should be given time off their regular work schedules so that they can be trained to use EMR systems effectively.

e. **Emphasize complementary process changes:** Managers and physicians should stop viewing EMR as a "mere technology." To derive value out of EMR, it has to be viewed as "Joint EMR-organization" endeavor. It is important to recognize that information technologies such as EMR interact with key organizational processes in order to bring about business value. Swanson (1994) calls such a joint endeavor as Type III innovation (Swanson, 1994). Complementary changes needs to be done to hospital processes such as patient registration, diagnosis, medical/surgical process, prescription generation and billing, and so forth.

f. **IT support:** Technical support is critical when implementing complex computer information systems. Problems can arise from hardware, software, network or user initiated failure. The provision of adequate IT support is essential to achieving maximum results of EMR with minimal downtime. Medical personnel also report feeling more comfortable transitioning to EMR when IT support is available. Health care institutions should never underestimate the value of IT support. During IT planning, institutions should consider acquiring IT support from EMR vendors and independent consultants, as well as establishing an in house support center.

CONCLUSION

Implementation of EMR systems can be a daunting task. Many health care advocates as well as the government are promoting EMR to become a standard in health care. The promised benefits of EMR are enticing. This process is also accelerated by changes in health care and reporting demands from the influx of managed care plans. We reviewed the existing literature as well as conducted our own study to unearth key barriers that affect EMR adoption. In this article, we highlighted these challenges and proposed solutions.

We urge researchers to continue this line of research. Several interesting topics still need to be explored. Is cost really the main inhibitor of EMR adoption? What organizational factors affect the success or failure of EMR adoption? Initial evidence suggests that the size of a practice matters with EMR; more work is needed to confirm this. What demographics data affect EMR adoption; for instance, does age or medical school education of physicians affect EMR adoption/use. The literature reports that many of the benefits of EMR will be realized later. As such, this gives us a need to conduct longitudinal studies at health care organizations.

Additionally, researchers should seek to evaluate EMR adoption research with the theoretical lenses of transaction cost economics, resource based view, social network theory, digital options theory, diffusion of innovation, and other grounded IT adoption theories.

REFERENCES

Adler, K. G., & Edsall, R. L. (2005). Electronic health records: A user-satisfaction survey. *Family Practice Management, 1*(1), 1-6.

Audet, A.-M., Doty, M., Peugh, J., Shamasdin, J., Zapert., K., & Schoenbaum, S. (2004). Information technologies: When will they make it into physicians black bags? *Medscape General Medicine, 6*(4), 2-7.

Baharozian, D. B. (2005). Electronic medical records in practice: Are we there yet? *Ophthalmology Times, 30*(22), 45-52.

Bar-Lev, S., & Harrison, M. I. (2006). Negotiating time scripts during implementation of an electronic medical record. *Health Care Management Review, 31*(1), 11-18.

Berkowitz, L. L. (1997). Breaking down the barriers: Improving physician buy-in of CPR systems. *Health care Informatics, 14*(10), 73-76.

Brynjolfsson, E., & Hitt, L. (2000). Computing productivity: Are computers pulling their weight? *Mimeo, MIT and Wharton, 1*(1), 21-43.

Burt, C. W., Hing, E., & Woodwell, D. (2005). *Electronic medical record use by office-based physicians.* United States. Retrieved February 14, 2008, from http://www.cdc.gov/nchs/products/pubs/pubd/hestats/electronic/electronic.htm

Devaraj, S., & Kohli, R. (2000). Information technology payoff in the health care industry: A longitudinal study. *Journal of Management Information Systems, 16*(14), 41-68.

Devaraj, S., & Kohli, R. (2003). Performance impacts of information technology: Is actual usage the missing link? *Management Science, 49*(3), 273-289.

Edsall, R. L., & Adler, K. G. (2005). An EHR user satisfaction survey: Advice from 408 family physicians. *Family Practice Management, 2*(1), 29-36.

Hough, C. B. H., Chen, J. C. H., & Lin, B. (2005). Virtual health/electronic medical record: Current status and perspective. *International Journal of Health care Technology and Management, 6*(3), 257 - 275.

Hsieh, C., & Lin, B. (2006). *Electronic medical records: current status and future development.* Paper presented at the National Decision Sciences Institute Conference Proceedings, San Antonio, Texas.

Lenhart, J., Loomis, G., Criswell, D., & Meggs, W. (2000). An analysis of trends, perceptions, and use patterns of electronic medical records among U.S. family practice residency programs. *Family Medicine, 32*(2), 109-114.

Miller, R. H., & Sim, I. (2004). Physicians use of electronic medical records: Barriers and solutions. *Health Affairs, 23*(2), 116-126.

Retchin, S., & Wenzel, R. (1999). Electronic medical record systems at academic health centers: Advantages and implementation issues. *Academic Medicine, 74*(5), 493-498.

Rind, D., & Safran, C. (1993). *Real and imagined barriers to an electronic medical record.* Paper presented at the Proceedings of the Annual Symposium of Computing Applications in Medical Care.

Simon, S. J., & Simon, S. J. (2006). An examination of the financial feasibility of electronic medical records (EMRs): A case study of tangible and intangible benefits. *International Journal of Electronic Health care, 2*(2), 185-200.

Swanson, E. B. (1994). Information systems innovation among organizations. *Management Science, 40*(9), 1069-1092.

Swartz, N. (2005). Electronic medical records' risks feared. *Information Management Journal.*

Valdes, I., Kibbe, D., Tolleson, G., Kunik, M., & Petersen, L. (2004). Barriers to proliferation of electronic medical record. *Informatics in Primary Care, 12*(1), 3-9.

Wade, M., & Hulland, J. (2004). Review: The resource based view and information systems research: Review, extension, and suggestions for future research. *MIS Quarterly, 28*(1), 107-142.

Whitten, J. L., & Bentley, L. D. (2007). *Systems analysis and design methods.* New York: McGraw Publications

Winn, D. (2002). EMR: Not a question of when but a question of how. *Group Practice Journal, 1*(1), 1-5.

KEY TERMS

Electronic Medical Records System: An interorganizational information system that captures the essential components of a patient's medical encounter with the medical provider, including storage and retrieval of subjective, objective patient information, assessment, and plans for patient care (Lenhart et al., 2000).

EMR Challenges: Any factor that would impede the adoption and use of EMR.

Health Care Information System: An arrangement of information technology, people, data, and processes that interact to gather, process, store, and disseminate health care information (adapted from Whitten & Bentley, 2007).

IS Adoption: The deployment of information systems within an organization.

Challenges, Systems, and Applications of Wireless and Mobile Telemedicine

Moshe A. Gadish
University of Guelph, Canada

Mieso K. Denko
University of Guelph, Canada

INTRODUCTION

Telemedicine can be defined as the delivery of healthcare, and the sharing of medical knowledge over a distance, using information and communication technologies (Pattichis, Kyriacou, Voskarides, Pattichis, Istepanian, & Schizas, 2002). Telemedicine aims at providing expert-based medical care to any place where healthcare is needed. It was introduced about three decades ago, where the only available communication technologies were telephone and fax machines. However, in recent years, many telemedicine applications have been developed over wired telecommunication technologies, wireless telecommunication technologies, as well as integrated wired and wireless networks.

Wireless communication and networking technologies such as 2.5G, 3G, 4G, and 802.11 Wireless LANs and Bluetooth have allowed the provision of wireless telemedicine systems, thus freeing the medical personnel and the patient from being bound to a fixed location. The wireless communication technology provides more flexible access to doctors and patients with the use of mobile stations. Mobile computers are gradually gaining more interest as the point of care tools used by many physicians and healthcare professionals. Since their introduction in the early 1990s, Personal Digital Assistants (PDAs) are increasingly becoming more and more popular for a large variety of medical applications. PDAs can be used in many areas of medicine as off-line reference tools, risk estimation for specific diseases, off-line diagnosis code databases, drug reference databases, and many more useful emerging applications. Some of these applications are stand-alone mobile applications, while many others are networked.

There are many benefits which wireless and mobile telemedicine can provide to healthcare and medical applications. Some of the specific benefits are outlined as follows. First, it can provide rapid responses to critical medical care, regardless of geographical barriers. It can be quickly deployed in disaster recovery areas, where existing communication links may have been disrupted. Second, it can provide flexible and reliable access to expert opinion and advice at the point of care with insignificant delay, and with improved management of medical resources. Third, it can allow patients to remain in their communities and receive medical services. This significantly reduces the costs of healthcare through improved healthcare management systems and reduced travel expenses. Fourth, it can provide interactive medical consultation and communication of medical records, image and video data in mobility scenarios, and with global coverage and connectivity, and fifth, it can support the empowerment and management of medical expertise in rural and underserved areas through the use of wireless infrastructureless networking technologies.

CHALLENGES IN MOBILE AND WIRELESS TELEMEDICINE

Mobile and wireless telemedicine faces a number of technical challenges emanating from various sources, including design and application scenarios. In this article, we will discuss the challenges related to image and video transmission, wireless signal quality, design, and standards. These challenges are briefly described in the following section.

Image and Video Transmissions

A telemedicine system would require the transmission of digital images as well as digital video. There are no set bandwidth requirements for the transmission of medical images. Some medical images, like a chest cardiogram, can take up to 50 MB of space, and would require longer transmission times. The transmission of medical images would require the use of compression algorithms. There are two types of medical image compression techniques: lossless, and lossy compressions (Kivijarvi, Ojala, Kaukoranta, Kuba, Nyul, & Nevalainen, 1998). Lossless techniques restructure the image exactly from its compressed format, preserving original quality, but providing limited compression ratios. On the other hand, lossy compression ratios provide an approximate reconstruction of the original image. The use of lossy compression in wireless telemedicine would require a careful evaluation of the effect of the compression on different medical diagnostic scenarios. When appropriate, the background of an image could be ignored, since it may be of no diagnostic value. Adapting such techniques can avoid compression of the background and other insignificant regions, providing substantial improvements of compression ratios (Young, Whiting, & Foos, 1999).

A reliable wireless telemedicine system also requires optimal transmission of digital video. In telemedicine, video transmission can be segregated into several categories, namely: real-time video transmission, off-line video transmission, medical video and audio for diagnostic applications, and nondiagnostic video and audio. Real-time diagnostic audio transmission includes transmission of stethoscope audio, as well as audio that accompanies diagnostic video (Pattichis et al., 2002).

There are a number of research projects investigating various techniques of video compressions. One possible technique for achieving acceptable video compression for diagnostic proposes the use of object-based encoding and decoding. Here, the video is segregated into different segments of diagnostic, and then assigned a corresponding bit rate, according to the predetermined significance. This technique significantly reduces the bandwidth, while retaining high-quality video images of regions with diagnostic interest (Pattichis et al., 2002). The difficulty of using object-based encoding is the uncertainty over which part of the video is diagnostically significant. Using interactive video may be enough in overcoming this uncertainty.

Many applications often transmit different types of streams simultaneously, such as medical images, vital signs, real time video, and other readings from various medical sensors. Depending on the application, the wireless transmission requirements for each stream can be quite different. Thus, a challenge when implementing a wireless telemedicine application would be to establish policies to coordinate, prioritize, and compress the various media streams to eliminate signal impedance and congestion problems.

Wireless Communication

The actual throughput of current commercially available cellular wireless links is fluctuant (Chu & Ganz, 2004). For example, CDMA provides data rates of about 153.6 kbps. However, the actual throughput of such cellular wireless links is fluctuant with an average throughput of about 70 kbps. Hence, the implementation of a mobile telemedicine system over such a link will have to take this limitation into account. Other forms of wireless communication technology also have their share of concerns. The use of satellite communications demands the procurement of expensive equipment, dedicated links, as well as skilled operators. Short range wireless technologies, like wireless local area network (WLAN) and Bluetooth, offer data rates that can support multimedia information—unfortunately, over short distances and with degrading results with increasing distance. The 802.11 WLANs offer data rate of 54 MBPS, while Bluetooth offers rates of up to 723 kbps. The 2G and 3G Cellular technologies offer the advantage of distance, but with lower maximum data rates. The maximum theoretical data rates for global system for mobile communications (GSM), general packet radio service (GPRS), EDGE, and universal mobile telecommunications system (UMTS) are 9.6 kbps, 171.2 kbps, 384 kbps, 2 mbps, respectively. The choice of the type of wireless communication technology to be employed depends on several factors, including deployment location, desired quality of transmission, and security level.

Design Challenges and Standards

One of the major challenges for the adoption of mobile data applications is getting the right form factor. This is also true in telemedicine, where it is significant to build the right device to handle the need of the healthcare service it provides. The challenge here is in getting the devices to a manageable size, thus minimizing inconvenience to the user (patient or the practitioner) (Tachkara, Wang, Istepanian, & Song, 2003). Patient confidentiality during use of a handheld device is an area of concern. Security should be factored into the design of the application, since the reliability and quality of patients' information that is stored are dependent on this. The most common solution is password protection. Automatic lockout after a user-defined period of time can also be used. There is also the problem of mapping free text notes into the handheld's understandable codes and structure. Physicians typically store their notes in computers via dictation and transcription. The translation of these notes into digital format by hand notes is costly. Tools to automate this translation are needed. Unfortunately, computer programs still cannot accurately interpret unconstrained text. The alternative would be having the physician coding his/her own data as they enter it through selection menus and other GUI controls (McDonald, 1997).

Each healthcare location may use a different primary computer system, and very likely different laboratory, pharmacy, and radiology service. Each location carries a part of that patient's medical information. Even within a single organization, like a hospital, many separate sources of information exist, containing different data formats and levels of granularity with different coding systems to identify similar clinical concepts. External healthcare sites may contain even more discrepancies than those within the hospital—employing different ways for representing patient data, drugs codes, and image formats. Such differences present information barriers to combining patient data from many sources in a single electronic medical record (EMR) system. The solution to this problem lies in the standards which have started to develop in the recent years (Tachkara et al., 2003), linking the differences between the many sources of electronic patient data, so that the data can be seamlessly combined and transferred with little cost. These standards allow vendors to create telemedicine systems, which can incorporate data from various sources like laboratories, pharmacy, and many more.

Some of the existing medical standards should be used when designing a mobile and wireless telemedicine system. The key standards are: Health Level 7 (HL7), Digital Imaging and Communications in Medicine (DICOM), Systematized Nomenclature of Medicine (SNOMED) and the International Classification of Diseases, Clinical Modification, Ninth Revision (ICD-9-CM) (Bidgood, Horii, Prior, & Van Syckle, 1997; McDonald, 2005; NCHS, 2000).

HL7 is the message standard of choice for communicating clinical information, like diagnostic results, notes, referrals, scheduling information, nursing, problems, clinical trial data, and so on, among heterogeneous computer systems. It is an international standard, in use in countries like the U.S., Canada, New Zealand, Japan, and many countries in Europe. HL7 provides the structure for interchanging patient information between source systems like laboratory, dictation, and pharmacy systems data repositories performance databases and medical record systems (Kurtz, 2002). DICOM is the standard of choice for transmitting diagnostic images. It is supported by all medical imaging vendors, and is tightly coupled to the HL7 standard (Bidgood et al., 1997). Universal code systems are now available for subject matter such as Logical Observation Identifiers Names and Codes (LOINC) for identifying laboratory observations. Other standards such as SNOMED exist for device classifications, organism names, symptoms, and pathology. ICD-9-CM is used in assigning codes to diagnoses associated with inpatient, outpatient, and physician office utilization in the U.S. It is maintained by the National Center for Health Statistics (NCHS) and the Centers for Medicare and Medicaid Services in the United States (NCHS, 2000).

All of the above standards are becoming more and more in use with various telemedicine systems, wired, wireless, mobile, or stationary. There are still a large number of systems in use that do not comply with them. The sooner every one of them adopts these standards, the faster and easier it will be to build interoperable telemedicine systems and applications.

MOBILE AND WIRELESS TELEMEDICINE SYSTEMS

A number of applications and systems have been proposed and developed for mobile and wireless telemedicine in recent years. They can be grouped according to the type of wireless technologies they employ, such as cellular or IP-based, and the type of services they provide, such as emergency healthcare, telecardiology, teleradiology, or remote monitoring. In this section, some existing mobile and wireless telemedicine systems are discussed.

Ambulance Emergency Services

Mobile telemedicine systems have been developed and tested for ambulance emergency services. Such a system consists of two main components: a mobile unit installed within the ambulance, and a base station for receiving hospital Internet connections (Tachkara et al., 2003). Using wireless digital cellular communications technologies, the mobile unit transmits to the trauma center real-time patient vital sign data, audio, and video images of the patient within the ambulance. It uses a video system with a patient monitoring interface to automatically capture images using a lossy JPEG image format, and then transmitted to the base station. The spectrum uses up to four bonded digital wireless phone lines to transmit voice, images, and data. In the hospital, a server receives these signals, and then forwards the transmitted data to a computer monitor to be viewed in an Internet browser setting by the doctor. One major limitation of this system is that it does not support real-time digital video. The video system is a movable camera mounted in the ceiling of the ambulance above the patient head, capturing images at 30 frames per second. In emergency situations, the system allows on-the-fly tradeoff between motion handling and resolution. There is a special button that lets the ambulance personnel capture video clips at 5 fps, which are sent in a store-and-forward fashion.

An ambulance system that supports real-time transmission of critical biosignals (electrocardiogram (ECG), blood pressure, heart rate, temperature, and SpO2) and still images of the patient has been in Europe (http://www.biomed.ntua.gr/emergency112/). The system utilizes a GSM link for transmissions. Originally, the system allows telediagnosis and teleconsultation with mobile healthcare providers by experts located at emergency coordination center or a specialized hospital. At a later point in time, the system was extended to provide a more integrated solution. This system, EMERGENCY-112, enables the transmission of critical biosignals and still images from the emergency site to an emergency call center. It is designed to operate over several communication links, including satellite, GSM, and Integrated Services Digital Network (ISDN). EMERGENCY-112 comprises of two modules: a patient unit operating automatically over several communication links, and a physicians unit located near the expert doctor in the hospital. The physicians unit can also operate over several communications links, depending on the expert's location. Given its ability of use of both wireless and wired communications, EMERGENCY-112 can be used to provide healthcare from ambulances, rural hospital centers, or any other remotely located health center, and even navigating ships.

Another ambulance system is the Teletrauma system (Chu & Ganz, 2004). The system is comprised of two units: a trauma-patient unit, and a hospital unit. The trauma-patient unit resides in an ambulance laptop or tablet PC. Through its connection to devises such as vital sign monitoring devices, portable ultrasound, and video camera, the unit collects patient's information, processes (i.e., compresses) it, and transmits it over 3G wireless link to the hospital unit PC in the trauma center, where it is processed and viewed. At the hospital, the physician has access to visual information like ECG, medical images and real-time video. Because the three media streams are transmitted simultaneously, congestion becomes a factor. To minimize the effect of congestion, the system effectively utilizes different transmission protocols to the different streams. The reliable transmission control protocol is used for the ECG and images to guarantee their transmission integrity. Due to video's tolerance to frame loss, its communication is managed using UDP, coupled with a frame rate adjustment policy.

Telemedicine in Space

National Aeronautics and Space Administration (NASA) has been at the forefront in the field of telemedicine, since the 1970s (Jones, McGinnis, Hamilton, D'Aunno, Simmons, Melton, & Armstrong, 2002). NASAs interest in telemedicine stems from the potential of using telemedicine in care of astronauts operating beyond Earth's orbit in the future. A telemedicine program developed by NASA comprises the astronaut patient, a consultant, data acquisition and handling hardware/software, and telecommunication connection. A device called Telemedicine Instrumentation Pack (TIP) collects medical audio, video, and data from the patient in space. TIP collects data like heart rate, blood oxygenation, blood pressure, and EKG, and it transmits medical video of eye, skin, ear, nose, and throat. The system contains an electronic stethoscope that collects heart, lung, and bowel sounds to ground flight surgeons. TIP data is output via the space stations or shuttle communication network back to earth.

Telecardiology System and Ultrasound

Telecardiology is concerned with diagnosis and treatment of diseases of the heart and circulatory system at a distance. One proposed system for telecardiology consists of a mobile van that houses a whole-body computed tomography (CT) scanner, and a second van that houses satellite communication equipment (Takizawa, Sone, Hanamura, & Asakura, 2001). This system features CT scanning, online two-way transfer of image data, and teleconferencing with a consultation center with various medical experts. The system personnel are the drivers, and a radiology expert that operates both CT scanner and the telecommunication unit. The van hosts a PC that manages CT data transmission teleconferencing equipment, an image printer, a facsimile machine, and resuscitation equipment. Communication between the mobile and stationary unit is done at 155 mbps in asynchronous transmission mode (ATP) via satellite or ISDN. Image communication uses a DICOM 3 protocol.

TeleCardio-FBC is a Web-based telemedicine system developed and deployed in Brazil that enables cardiologists at the unit of cardiology and cardiovascular surgery to cooperate with other physicians (Monotoni et al., 2002). The system provides cardiology care for patients living in remote areas, reducing costs and allowing better follow-up services for discharged patients. This system was designed for use with desktop computers as the only computational platform. Telecardio Mobile was later developed with the purpose of taking full advantage of 3G technology, adding mobility to the original system. It provides online access to the information in the TeleCardio-FBC system through PDAs and mobile phones (Monotoni, Villela, Rocha & Rabelo, 2002). This system consists of two subsystems: M-TeleCardio, and WapCardio. M-TeleCardio allows access to full functionality offered by TeleCardio-FBC to the mobile device (Laptop, PDA, palm, mobile phone) that is wirelessly connected to the Internet. WapCardio subsystem utilizes wireless application protocol (WAP) technology to supply medical information like remote consultation requests and results of medical procedures to physicians on mobile.

Adding a wireless capability to ultrasound would enable image scans directly from the patient rooms, and immediate doctor consultation from distant locations. The purpose of a wireless ultrasound application is to create an electronic archival system for ultrasound images for retrieval through wireless medium at remote locations (Kazmi, Arslan, Tulgar, Pedersen, Pahlavan, Cengiz, & Arslan, 2000). The doctors will take an ultrasound scan of the patient, save it on the wireless laptop, and then transmit it to a centralized server for future references. The application also includes an option for the doctor to annotate voice and text files to the ultrasound image file. Unique to the wireless ultrasound application is that it can handle three different file formats, including image, voice, and text, which are all transmitted and received as a single object between the laptop and the server archive.

PDA-BASED MOBILE AND WIRELESS TELEMEDICINE APPLICATIONS

PDAs are increasingly becoming more popular in use within the field of medicine. Physicians are showing interest in computer-based information about medications, patient education material, and treatment recommendations. Such interest has motivated a lot of software vendors in developing various types of medical applications for the PDA platform (Fisher & Steward, 2003). Current PDAs do not have the functionality to manage a complete electronic medical record system, nor are they capable of storing large graphics. Regardless of these limitations, they have been identified by physicians as excellent tools for managing clinical information and accessing it at the point of care (Bates, Ebell, Gotlieb, Zapp, & Mullins, 2003). This section presents an overview of current literature regarding the applications of handheld devices in medicine.

Basic PDA Applications

A PDA comes with many basic functions that are built into the operating system. These functions can be quite valuable for the medical practice and healthcare. Some of these include: scheduler, address book, a to-do-list, and memo functionality. A physician uses the scheduler to set up alarms alerting him of upcoming appointments with patients or other medical practitioners. The latest version of Windows CE5 operating system comes packaged with such basic applications (Ebell, Gasper, & Khurana, 1997).

Another basic functionality is the ability to read text files, such as medical books, references, and documents downloaded from the Internet. Several vendors offer such reader applications for PDA platform. Adobe has a PDA version of its reader, while AportisDoc Mobile and TealDoc (http://www.tealpoint.com/softdoc.htm) are two other document reader applications.

Voice dictation applications have been proven to be very useful for the physician. As the physician speaks, the dictation software automatically writes it as text, that is saved as a text file that can be retrieved later for analysis by the same doctor or another. Many PDAs come equipped with voice recognition or dictation software that can be used for this purpose.

Medical Information Access

It is important for the doctor at the point of care to have access to medical information. Access to medical information is becoming more important when considering the rate and amount of newly added medical information, and the increase in expectations to follow medical guidelines. There are several applications allowing access to internet-based literature. AvantGo is a free program that allows the mobile user to download Web sites on his handheld PDA on a daily basis for off-line viewing (http://www.avantgo.com). During synchronization, medical journal Web sites that the doctor subscribes to are downloaded into it (or updated if the content is out of date). This allows the doctor to have up-to-date abstracts as well as full articles in relevant medical journal on their PDA, at any time.

Drug information databases and pharmacopoeias can be formatted onto a handheld device. Studies have shown that drug applications can save physicians valuable time during information retrieval, and that it easily incorporates into their workflow, making improvements in drug-related decision making. Inquiries about drugs at the point of care provide prescription information at the patient's bedside, potentially reducing the information search time, as well as mitigating the risk of incorrectly prescribing a harmful drug (Rothschild, Lee, Bae, & Bates, 2002).

There are several mobile applications for handheld devices. A very popular handheld drug database is ePocrates. This free medical reference guide provides a comprehensive drug list, dosing guidelines, drug interaction guidelines, retail pricing of drugs, and common side effects. Another popular free database, the ABX Guide (http://hopkins-abxguide.org/show_pages.cfm?content=siteinfo_content.html), was developed at the division of Infectious Disease at the John Hopkins University. It allows searches for antibiotic information with respect to diagnosis, pathogen, or antibiotic type. The database is always up to date, and it can be transferred to a Palm-based handheld to be used at the point of care. Another mobile drug reference system is the *mobilePDR* (mobile Physician's Desk Reference) (Bojovic & Bojic, 2005). The *mobilePDR* consists of several components: an application and a corresponding database on a PDA device, desktop software for backing up data and synchronization over the Internet with Medec Web server that delivers updates to drug data.

Patient Tracking Applications

PDAs also play a very significant role in areas where patient tracking is needed. Handheld devices can be used to monitor patients' health, and to keep medical staff informed of his/her condition. Recently, many patient tracking programs have become available. Patient Track, which is available for both Palm OS and Windows CE PDAs, allows the user to enter patient records, including demographics, medication/allergy lists, test results, and radiology reports. Another application of interest is the Ward Watch (Palm OS). This application aids medical staff involved in ward rounds by recording investigations, medications, dosages, and consultations (Duncan & Shabot, 2000; Fischer & Steward, 2003). The TelePatient System assists healthcare providers in monitoring patient's information at all times (Chu & Ganz, 2004). The system involves two networks: a wireless personal area network that contains the medical data acquisition units, and a wireless network for transmitting the data to the healthcare provider. Here, a PDA is used as a gateway between the two networks. Medical data acquired by the various medical acquisition units (i.e., cardiac monitor, camera) is passed through the PDA to the healthcare provider.

There are risks in the areas of patient data management and sign-over between physicians, with respect to propagation of errors, when done by hand using pen and paper means. PDAs may play a paramount role in mitigating these human errors. Documents containing patient information can be transferred from the desktop computer to the PDA or created on the PDA. Changes to the data made on the database can be transferred to the PDA during synchronization. Similarly, data manipulated on the PDA can be sent to the desktop application or database in order to update its out-of-date information. Infrared beaming or Bluetooth communication can be used to transfer information between the PDAs of different physicians, allowing them to exchange patient information between shifts.

Medical Research and Education Applications

The medical education environment is using PDAs for electronic data collection to manage information. Various medical education institutes are utilizing PDAs for the evaluation of training of medical students, and in the monitoring of student's clinical experience. The University of Minnesota Medical School has equipped its students with PDAs loaded with clinical references and tools to collect information about the patients they see and the seminars they attend during their clerkship (Speedie, Pacala, Vercellotti, Harris, & Zhou, 2001). A survey to program directors of all U.S. Medical Schools found that two thirds of the schools were using handheld devices in their residency programs (Fischer & Steward, 2003). A PDA application that logs and tracks resident's procedural experience has been shown to be a useful tool for ensuring adequate educational experience.

PDAs outfitted with tailored application for data collection have been used for research data collection. A case in point is the use of a PDA for the evaluation of craving and withdrawal symptoms of patients in a smoking cessation program. Instead of asking the patients to remember their experiences looking back, they can enter this information in real-time at their natural environment. This type of point of care data collection, using a handheld device, has been shown to be quite efficient—23% faster than traditional hand recording (Fischer & Steward, 2003). The size advantage of a PDA, being small and handy, is that it can easily be carried in the briefcase or a pocket, and data entry can be carried out at any time.

Business Applications

PDAs can play an important role in increasing medical business efficiency. Business processes are improved by utilizing PDA's basic applications like schedulers, to-do-lists, and remainders. Another way to improve business efficiency is by handling the billing process. Billing information can be entered to the PDA by the physician during rounds in the hospital, and then transferred to a central database (later on during synchronization) and viewed by office staff. This approach provides quicker billing and improves the business cash flow.

Prescription and Pharmaceutical Applications

Using a handheld computer to do electronic prescription is becoming increasingly popular. Many physicians are now writing prescriptions electronically and transmitting them to the pharmacy without manual intervention. Using this approach mitigates the number of errors that may have been otherwise introduced by illegible handwriting or dictation, as well as saving time for the physician, pharmacist, and patient (Embi, 2001; Fischer & Steward, 2003).

FUTURE TRENDS IN MOBILE AND WIRELESS TELEMEDICINE

The next generation mobile and wireless communication technologies coupled with heterogeneous network tremendously improve the future mobile and wireless telemedicine services. There is an increasing research for a mobile telemedicine system that can rapidly be deployed in remote areas, or in disaster recovery situations where telecommunication infrastructure is lacking or destroyed. One such system is the Tele-Emergency system, which is based on mobile ad hoc networks (MANETs) (Husni, Heryadi, Woon, Arifianto, Viswacheda, & Barukang, 2006).The basic requirements for such systems are the capability to work in remote areas with little or no existing communication infrastructure, and provide both basic and extensive medical services. The system is designed to manage medical record data, as well as deliver real-time multimedia and geographical information system data wirelessly. In MANETs, mobile nodes communicate directly with one another in a peer-to-peer manner, where each mobile node also acts as a router for any other node. The healthcare practitioners will use mobile nodes to transmit patient data to the healthcare center or mobile ambulance. The mobile ambulance is a mobile node equipped with a local server and optional communication devices. Every mobile node within the subnetwork can route information through a MANET routing protocol. Mobile and wireless telemedicine can also be deployed in heterogeneous networking environment, which includes wireless mesh networks (with infrastructure or hybrid architectures). Recent advances in wireless mesh networking that allows multiradio and multichannel transmission will significantly improve the performance of the future mobile and wireless telemedicine.

CONCLUSION

Wireless technologies can provide promising solutions in the future healthcare and medical applications. Current applications vary with respect to the wireless technology they employ and the type of service they provide. Whether they utilize 3G, 4G, or WLANs communication technologies, their wireless capability adds a much-needed dimension to existing wired and stationary healthcare services that result in the fast and successful delivery of healthcare services in remote areas. Mobility, coupled with wireless connectivity, has the advantage of providing more flexible healthcare services at the point of care. In many medical situations, the ability to provide such services saves lives in danger, or significantly mitigates the severity of medical conditions by offering a quicker alternative. Moreover, the use of PDAs in telemedicine can provide basic services like scheduling and medical billing. It provides a convenient way to wirelessly retrieve medical information on the go, and to make information immediately available at the palm of the physician's hand. The existing systems and applications can be further improved while the new ones emerge, making use of the advances in medical, communication, and networking technologies.

REFERENCES

Bates, D.W., Ebell, M., Gotlieb, E., Zapp, J., & Mullins, H.C. (2003). A proposal for electronic medical records in U.S. primary care. *Journal of the American Medical Informatics Association, 10*(1), 1–10.

Bidgood, W.D. Jr., Horii, S.C., Prior, F.W., & Van Syckle, D.E. (1997). Understanding and using DICOM, the data interchange standard for biomedical imaging. *Journal of the American Medical Informatics Association, 4*(3),199–212.

Bojovic, M., & Bojic, D. (2005). MobilePDR: A mobile medical information system featuring update via Internet. *IEEE Transactions on Information Technology in Biomedicine, 9*(1), 1–3.

Chu, Y., & Ganz, A. (2004). Mobile telemedicine systems using 3G wireless networks. *EEE Trans. Inf. Technol. Biomed., 8*(4), 456–462.

Duncan, R.G., & Shabot, M.M. (2000). Secure remote access to a clinical data repository using a wireless personal digital assistant (PDA). In *Proceedings of AMIA Symposium* (pp. 210–214).

Ebell, M.H., Gasper, D.L., & Khurana, S. (1997). Family physician's preference for computerized decision-support hardware and software. *J Fam Prac, 45*(2), 127–128.

Embi, P.J. (2001). Information at hand: Using handheld computers in medicine. *Cleve Clin J Med., 68*(10),840–853.

Fischer, S., & Steward, E.T. (2003). Handheld computing in medicine. *Journal of the American Medical Informatics Association, 10*(2), 139–149.

Husni, E.M., Heryadi, Y., Woon, W.T.H., Arifianto, M.S., Viswacheda, D.V., & Barukang, L. (2006). Mobile ad hoc network and mobile IP for future mobile telemedicine system. *IFIP International Conference on Wireless and Optical Communications Networks.*

Jones, J.A., McGinnis, P.J., Hamilton, D.R., D'Aunno, D., Simmons, S., Melton, S., & Armstrong, C. (2002). Isolation: NASA experiments in closed-environment living - advanced human life support enclosed system, final report. *American Astronautical Society*, 397–406.

Kazmi, A., Arslan, B.R., Tulgar, M., Pedersen, P.C., Pahlavan, K., Cengiz, B., & Arslan, A. (2000). Advances in wireless telemedicine: Improvements and challenges facing wireless ultrasound. Information technology applications in biomedicine. In *Proceedings of 2000 IEEE EMBS International Conference* (pp. 11–16).

Kivijarvi, J., Ojala, T., Kaukoranta, T., Kuba, A., Nyul, L., & Nevalainen, O. (1998). A comparison of lossless compression methods for medical images. *Comput. Med. Imag. Graphics, 22*(4), 323–339.

Kurtz, M. (2002). HL7 version 3.0: A preview for CIOs, managers, and programmers. *Journal of Healthcare Information Management, 16*(4), 22–3.

McDonald, C.J. (1997). The barriers to electronic medical record systems and how to overcome them. *Journal of the American Medical Informatics Association, 4*(3), 213–221.

Monotoni, M., Villela, K., Rocha, A.R., Rabelo, A. (2002). TeleCardio mobile: Development of platform-independent telemedicine applications. In *Proceedings of Second Conference on Mobile Computing* (pp. 119–130). Heidelberg, Germany.

Pattichis, C.S., Kyriacou, E., Voskarides, S., Pattichis, M.S., Istepanian, R., & Schizas, C.N. (2002). Wireless telemedicine systems, an overview. *IEEE Antenna's and Propagation Magazine, 44,* 143–153.

Rothschild, J.M., Lee, T.H., Bae, T., & Bates, D.W. (2002). Clinician use of a palmtop drug reference guide. *Journal of the American Medical Informatics Association, 9*(3), 223–229.

Speedie, S., Pacala, J., Vercellotti, G., Harris, I., & Zhou, X. (2001). PDA support for outpatient clinical clerkships: Mobile computing for medical education. In *Proceedings of AMIA Symposium* (pp. 632–636).

Tachakra, S., Wang, X.H., Istepanian, R.S., & Song, Y.H. (2003). Mobile e-health: The unwired evolution of telemedicine. *Telemedicine and E-Health Journal, 9*(3), 247–257.

Takizawa, M., Sone, S., Hanamura, K., & Asakura, K. (2001). Telemedicine system using computed tomography van of high-speed telecommunication vehicle. *IEEE Transaction on Information Technology in Biomedicine, 5*(1), 2–9.

Young, S.S., Whiting, B.R., & Foos, D.H. (1999). Statistically lossless image compression for CR and DR. In *Proceedings of SPIE: Medical Imaging* (Vol. 3658, pp. 406–419).

KEY TERMS

Bluetooth: Wireless technology that enables data connections between electronic devices—such as desktop computers, wireless phones, electronic organizers, and printers—within a short communication range, typically about 10 meters.

Electrocardiogram (ECG): It measures heart's electrical activity.

General Packet Radio Service (GPRS): It provides mobile and wireless data services.

Digital Imaging and Communications in Medicine (DICOM): A set of standards for handling, storing, and transmitting information in medical imaging.

Health Level 7 (HL7): It is a standard for healthcare, and is the interface standard for communication between various systems employed in the medical community.

Mobile Ad Hoc Network (MANET): It is a self-configuring network of mobile nodes connected by wireless links.

Personal Digital Assistant (PDA): A hand-held computer device which manages personal information and interacts with other computers and telecommunications systems.

SpO2: A measurement of the amount of oxygen attached to the hemoglobin cell in the circulatory system.

Wireless Application Protocol (WAP): It is an application that allows users to access information instantly via handheld wireless devices such as mobile phones, pagers, two-way radios, smart phones, and PDAs.

Wireless Local Area Network (WLAN): WLANs are defined in the IEEE 802.11 standard.

Classification of Waste in Hospitals

Victoria Hanna
University of Melbourne, Australia

Kannan Sethuraman
Melbourne Business School, Australia

INTRODUCTION

The health care industry today is a complex web of ever-changing relationships among patients, physicians, hospitals, insurers, employers, communities, and government. A combination of factors, including the emergence of intense, dynamic competition (and consolidation) and increasing expectations of demanding patients, has generated a challenging environment for hospitals. Today, spending on health care in the United States has risen to $1.3 trillion, which almost equals 14.5% of the United States GDP. The cost of providing health care in OECD countries ranges from 7% to 9% of GDP (Folland, Goodman & Stano, 1997) and is expected to rise due to the aging population, an increasing number of chronic health conditions, soaring drug costs, and costs of new technology. With health care costs continuing to rise faster than general inflation, a theme that resonates throughout the health care field today and receives increasing attention from policymakers, academia, and industry is the necessity to contain costs without compromising quality of care. Here we interpret the quality of care as how cost effectively the hospital organizes its resources to meet the medical requirements of its patients.

Hospitals have to focus their efforts on identifying and eliminating waste of all forms if they are to succeed in today's competitive landscape. A recent study by the Murphy Leadership Institute (Murphy, 2003) concluded that wasteful work consumes more than 35% of hospital employees' time. This wasteful work includes activities such as completing multiple forms for the same task, filing inefficient shift-to-shift departmental reports, waiting for medications, and searching for misplaced records. Jimmerson warns that the actual amount of waste in health care lies closer to 60% (Panchek, 2003).

In this chapter, we briefly review principles of lean philosophy for improving performance and then present a classification of waste that is relevant to hospital management. This classification is aimed at directing hospital initiatives toward understanding and controlling waste in its health care delivery processes. Through several examples from real-life hospital case studies that we have investigated, we trace much of the waste to various types of variability (both natural and artificial) and offer prescriptions to control variability. We then provide some guidelines for streamlining processes and show how this would benefit various stakeholders. We conclude the chapter with some directions for further research.

BACKGROUND

Since the 1980s, hospitals have borrowed concepts and ideas that have helped transform manufacturing industries for decades. These include ideas from total quality management, lean thinking, and six sigma approaches. Although there are some overlaps among these principles, in this chapter, we focus primarily on lean principles and their applications in hospitals.

Lean Thinking in Hospitals

Toyota Production System principles, popularized by Womack, Jones, and Roos (1990) as lean production concepts, have helped turn around many manufacturing firms. Although Womack and Jones (1996b) state that the principles of lean production can be applied equally in every industry around the globe, the adoption of lean principles has largely remained limited to the manufacturing industries, and there is sparse evidence to suggest the effective "crossover" of these principles to nonmanufacturing settings. However, some recent research holds promise for the transferability of these principles to the service sector; in particular, to the health care sector. Lean patient care is all about creating more value for patients through the elimination of all nonvalue-adding steps in the health care delivery process. Wysocki (2004), Miller (2005), and Weber, Jimmerson, and Sobek (2004) have documented initiatives highlighting how lean thinking is helping to transform hospitals around United States.

In the following sections, we investigate how waste manifests in hospitals and offer prescriptions for its reduction/elimination.

MANIFESTATION OF WASTE IN HOSPITALS: CLASSIFICATION, REASONS, AND CURES

Drawing on the vast literature that exists on the Toyota Production System (Liker, 2004; Monden, 1993; Womack & Jones, 1996a; Womack et al., 1990), we offer a classification of waste, highlighting how waste propagates in hospitals. This classification is expected to assist both hospitals and other players in the health care value chain in their pursuit of the elimination of waste.

Waste Classification

Ohno (1988) identified and popularized the notion of seven wastes in manufacturing that include overproduction, waiting, unnecessary transport or conveyance, over-processing or incorrect processing, excess inventory, unnecessary movement, and defects. We provide a revised classification of waste that adapts each waste in Ohno's classification to fit the health care context.

1. **Over-servicing.** This waste identifies situations in which patients are being processed or served at a stage earlier or faster than their actual needs in subsequent stages. This is a symptom of patients being pushed through the system. Such an approach appears logical and cost-effective when batching economies or other constraints due to poor coordination are present. However, this push approach often creates greater congestion and larger queues. For instance, in many of the hospitals we have studied, patients scheduled to have surgery in the morning session in a theater were all requested to arrive between 6.45 and 7.15 A.M. Our investigation found that this practice was followed to allow anesthetists to complete their preassessment of patients prior to the commencement of surgery on the first patient. This practice is largely motivated by their need to be physically present in the theater for the complete duration of surgery, poor information transfer of patient medical records between surgeons and anesthetists, and the pressure from surgeons to ensure fast turnaround between surgical cases. As a consequence, there is a glut of patients undergoing preoperative assessment much ahead of their scheduled need, overburdening the subsequent stages resulting in chaotic movement of patients, staff, and information and long wait times.
   Another example of over-servicing is when diagnostic tests are performed much ahead of their actual requirement, causing redundancies and wastage of resources. In addition, requesting more diagnostic tests than what is required for accurate assessment, is an example of over-servicing patients.
2. **Waiting.** In hospitals, one can observe two kinds of waits—one experienced by patients and another by staff and doctors in the system. Under lean philosophy, both of these waits are categorized as waste. Patient waits correspond to poor service and are undesirable. Resource waits result in reduced utilization and increased costs for the hospital. Hospital practices focus on minimizing the wait for doctors and its staff and resort to batching strategies, which result in higher waits for patients. Lack of coordination across stages also results in excessive patient waits and doctor waits. Patient waits have other costs that are normally not well recognized by hospitals. For instance, you need more room to hold them, more resources to engage and monitor them, and more resources to progress them through the system.
   To improve performance on this measure, hospitals need to understand the implications of queuing and the causes and costs of congestion. Little's law from queuing theory provides a useful framework, illustrating the linkage between average patient wait time, the average number of patients in progress, and patient throughput rate.
3. **Unnecessary transport or conveyance.** Historically, hospitals have been designed by specialty rather than around the patient. As a result, we often find patients who are in the middle of their treatments moved a long distance, creating

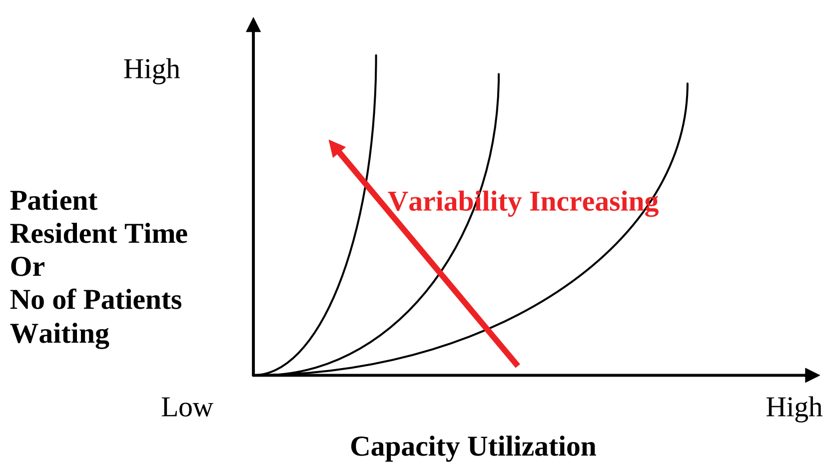
unnecessary transport. Also, they are moved in and out of waiting bays or postoperative wards to accommodate capacity restrictions elsewhere. To prevent such redundant transport, a coordinated approach to patient process flow is a must. Specialty hospitals such as Shouldice Hospital (Heskett, 1983) in Toronto have streamlined their process flows around patient needs, minimizing movement of patients through the system.

We have also witnessed a convoluted flow of documents (between doctors and hospitals), medication, and other shared ancillary equipment and staff, which often result in a waste that can be categorized as unnecessary transport.

4. **Unnecessary movement.** Unnecessary motion is caused by poor workflow, poor layout, and inconsistent or undocumented work methods. Any wasted motion that nurses, surgeons, or orderlies have to perform during the course of their work, such as looking for reports, reaching for reports or patients, can be classified as unnecessary movement. Also, any unnecessary walking that they perform will fall into this category.
5. **Over-processing or incorrect processing:** In hospitals, processes evolve over time and are typically not optimized in a holistic manner. Each specialty focuses its effort on enhancing patient care from its own perspective. As a result, you have unnecessary or over-processing and sometimes incorrect processing. Unnecessary surgeries and diagnostic tests are well-known examples of this type of waste. Rother and Shook (1999) provide a technique called value stream mapping that enables hospital officials to document and understand the steps that are essential for serving their patients more effectively. From a clinical perspective, the development of pathways has enhanced the quality of care and has facilitated earlier detection of deviation from expected treatment paths.
6. **Excess Inventory.** This waste is easily identifiable in hospitals. There is plenty of evidence for excess stock of medical supplies, pharmacy items, and other medical equipment resulting from uncoordinated purchasing policies.

In a hospital environment, patients in progress (PIP) also represent a form of inventory, and the level of PIP is determined by batching considerations and variability. Litvak and Long (2000) identified four forms of variability that occur in health care systems; namely, clinical, flow, professional, and artificial variabilities. Clinical variability captures the degree of illness, choice of treatment alternatives, and response that patients with the same disease exhibit. The flow variability is illustrative of the randomness inherent in patient arrival rates, and professional variability illustrates the heterogeneity in the ability of medical practitioners and health care delivery systems to provide the best treatment. The fourth form of variability was branded as artificial, and is an artifact of dysfunctional management and inefficient health care delivery practices. These variabilities increase PIP and the average patient resident time in the system. The relationship between PIP or patient resident time and variability

Figure 1. The ill effects of variability and capacity utilization



is conceptually shown in Figure 1. Based on this relationship, we can conclude:

i. Increased variability (either in the patient arrival process, the patient service process, or both) will essentially cause patient queue length and average patient resident time to increase.
ii. In a highly utilized environment, increased variability will result in the patient queue length and average patient resident times increasing in a highly nonlinear fashion.
iii. With higher levels of variability, the system will have to operate at lower levels of utilization to achieve target responsiveness (measured in terms of patient resident times).

It is critical for hospitals to get a better handle on various types of variabilities that cause havoc on patient responsiveness. Clinical and professional variabilities are virtually impossible to manage; however, flow variability can be managed though improved scheduling practices that attempt to level out the workload. Artificial variability arising from dysfunctional practices has to be minimized if not eradicated completely.

7. **Defects.** Medical service errors such as wrong medication or wrong procedures performed on a patient constitute defects in a hospital context. These errors can be life threatening, and the costs of such errors should not be measured only in terms of dollar value. They also waste resources in at least four ways. First, the materials (e.g., medical supplies and other inputs) that were consumed to provide the service cannot be recovered; second, the labor (e.g., doctor and other personnel time) used to provide this service is wasted; third, you require labor to perform the rework; and last, labor is required to address any forthcoming patient complaints (e.g., litigation, etc.) and word-of-mouth adverse publicity. Some medical errors occur because the concept of error proofing is less prevalent in health care, and due to problems not being tackled at their source. Unless hospitals build a culture of empowering its workforce to stop and fix problems at their source, it is very hard to eliminate this type of waste. For example, Wysocki (2004) emphasizes the need to empower workforce through an example of a nurse who was empowered to bring immediate attention to a situation that otherwise might have been life threatening to a patient.

   In a recent book, Liker (2004) extended Ohno's waste classification to include an eighth waste—unused employee creativity.

8. **Unused employee creativity.** This refers to the time, ideas, skills, improvements, and learning opportunities that a hospital foregoes by not engaging or listening to their employees. People who are closest to their processes know the finer details of the processes and how they can be improved. Few hospitals have tapped this resource meaningfully.

Waste Minimization Strategies

In this section, we prescribe approaches that would help minimize the occurrence of waste in hospitals. See Table 1 for a summary.

1. **Standardize work procedures and reduce variability:** Standardization of work procedures would help minimize artificial variability created through dysfunctional processes and facilitate building in quality. If the errors persist in spite of following standardized work, then it is important to continuously strive to improve the standards.
2. **Streamline patient flow:** The smooth flow across stages of the treatment process would ensure that processes are connected meaningfully with one another without any loss of information from one stage to another. A tightly coupled process would help highlight bottlenecks and imbalances in the system more effectively. As Minoura, a former president of Toyota Motor Manufacturing, North America, mentioned (Liker, 2004), problems in single-piece flow manufacturing forces everyone to solve the problem immediately and enables everyone to become better team players. Similarly, coordinated patient flow has the potential to not only build in quality but also improve the responsiveness of the system.
3. **Improve coordination across stages in the health care value chain:** It is not only important that the hospital manages the value creation process within its boundaries effectively, but it is also critical to coordinate its interface with external parties involved in the treatment process. When the information flow is disrupted between

Table 1. The causes, contributing factors, and countermeasures for different types of waste in the health care industry

Types of Waste	Contributing Factors	Countermeasures
1 Over-servicing	Batching economies Poor coordination and information sharing Lack of goal congruence	Modify the measurement and evaluation systems to facilitate the use of systemwide objectives Adopt pull approach as opposed to the push approach to manage patients through the system Have a common database that can provide real-time updated information to all concerned parties
2 Waiting	Floating bottlenecks Imbalance in capacity across stages Variability in patient arrival and service processes Unexpected disruptions Lack of information sharing across different stakeholders	Minimize the variability in patient mix and schedule in elective surgeries Keep buffer capacity during instances of increased uncertainty as in emergency cases Align incentives to improve data sharing
3 Unnecessary transport or conveyance	Poor facility layout Lack of real-time updating of patient information	Design layouts around patient flow when possible Ensure centralized real-time updating of information
4 Unnecessary movement	Undocumented or poor work methods Inefficient design of workflow Poor layout of facilities	Apply value stream mapping to eliminate nonvalue-added activities
5 Over-processing	Inefficient design of work flow Undocumented procedures	Standardize work methods and procedures
6 Excess Inventory	Batching economies Bulk discounts Imbalances in capacities across stages Uncertainty in supply Variability in patient service process Artificial variability introduced through dysfunctional practices	Develop strategic partnerships between surgeons, hospitals, and other stakeholders Initiate vendor-managed inventory for medical suppliers where possible Initiate collaborative planning and forecasting methods for managing demand
7 Defects	Lack of error proofing Batching and demand surges create a pressurized environment, leading to increased errors Lack of worker empowerment to fix problems at source	Empower employees and ensure quality at the source at all opportunities Develop KPIs to reflect strategic priorities
8 Unused employee creativity	Lack of appreciation for human potential to innovate	Provide forums to give employees opportunities to share their ideas for improvement

the surgeons and the hospital staff, it can create multiple opportunities for propagation of waste through the system. Hence, all efforts need to be made to ensure smooth flow of real-time information across the different entities.

4. **Align incentives:** While the effective treatment of patients is the overarching objective, the priorities of various parties involved in the treatment process can be divergent, and meaningful incentives will help align the priorities.
5. **Adoption of consistent information technology:** In today's rapidly changing technological world, it is imperative for hospitals to invest in information technology that can take advantage

of centralized information. This would minimize duplication of efforts and clerical errors and achieve higher levels of data consistency and improved responsiveness.

6. **Build quality at the source:** Jidoka (building quality at the source) will ensure that the errors occurring at a particular stage do not flow downstream unattended and thereby cause even more waste or distress. Also, it facilitates the identification and rectification of root causes right at the source.
7. **Level out the workload:** Although it sounds harder to level out a work schedule in the service setting, it is not far fetched to see how this concept can be applied in hospitals. This is easier to do in case of activities that can be planned ahead of time. The key principle here is to minimize the unevenness in workload that may result due to an unleveled schedule. Also, Sethuraman and Tirupati (2005) have recently documented the evidence of the bullwhip effect in the health care sector. If the workload is not leveled out, then it has the potential to amplify the demand variation as one travels up the health care value chain.
8. **Use visual control to highlight problems:** Fujio Cho, president of Toyota Motor Corporation (Liker, 2004), said, "Mr. Ohno said one must clean up everything so you can see problems. He would complain if he could not look and see and tell if there is a problem." Visual control aids in noting whether the work is being performed in accordance with the standard, and it highlights any deviation from the standard. This will be very much applicable in case of hospitals, and it would refer to the design of visual cues and communication that enable fast and proper execution of operations and processes.
9. **Empower workforce:** People closest to the process have the most creative ideas to solve problems or improve the process. Hence, it is vital for hospitals to harness this resource meaningfully as manufacturers have done in the past.

CONCLUDING REMARKS AND DIRECTION OF FUTURE RESEARCH

The taxonomy of waste we have provided should aid hospitals and hospital management in their quest to eradicate waste from their business processes. However, the question still remains whether it is truly possible to translate concepts that have benefited the manufacturing sector into health care. Consider "single patient flow" as akin to single piece flow in the case of manufacturing—is it ever achievable in a hospital setting, and if so, would it deliver the same benefits that it has to manufacturing organizations, such as the elimination of waste, achievement of higher quality, improved safety and better productivity? More research is required to fully understand the benefits and consequences of translating useful principles from manufacturing into a language and form that the health care sector can easily understand and apply.

REFERENCES

Folland, S., Goodman, A.C., & Stano, M. (1997). *The economics of health and healthcare*. Upper Saddle River, NJ: Prentice Hall.

Heskett, J.L. (1983). *Shouldice Hospital Ltd*. Harvard Business School Case No. 9-683-068. Harvard Business School. Boston, MA. Available at www.hbsp.harvard.edu

Liker, J.K. (2004). *The Toyota way: 14 management principles from the world's greatest manufacturer*. New York: McGraw-Hill.

Litvak, E., & Long, M. (2000). Cost and quality under managed care: Irreconcilable differences? *The American Journal of Managed Care, 6*(3), 305–312.

Miller, D. (2005). *Going lean in health care*. Cambridge, MA: Institute for Healthcare Improvement.

Monden, Y. (1993). *Toyota management system*. Portland, OR: Productivity Press.

Murphy, M., (2003). *Eliminating wasteful work in hospitals improves margin, quality and culture*. Murphy Leadership Institute Research Briefing, No. 4. Retrieved from http://www.cs.odu.edu/~cpi/cpi-s2004/statag/murphy-wastefulwork.pdf

Ohno, T. (1988). *Toyota production system: Beyond large-scale production*. Portland, OR: Productivity Press.

Pancheck, P. (2003). Lean health care? It works! *Industry Week Magazine, 252*(11), 34–40.

Rother, M., & Shook, J. (1999). *Learning to see*. Massachusetts: The Lean Enterprise Institute.

Sethuraman, K., & Tirupati, D. (2005). Evidence of bullwhip effect in healthcare sector: Causes, consequences, and cures. *International Journal of Services and Operations Management, 1*(4) 372–394.

Weber, D., Jimmerson, C., & Sobek, D.K. (2004). Reducing waste and errors: Piloting lean principles at IHC. Submitted to the Joint Commission on Quality and Safety. Retrieved from http://www.coe.montana.edu/IE/faculty/sobek/IOC_Grant/JCJQS_submit_rev2.pdf

Womack, J.P., & Jones, D.T. (1996a). Beyond Toyota: How to root out waste and pursue perfection. *Harvard Business Review*, September-October, 140–153.

Womack, J.P., & Jones, D.T. (1996b). *Lean thinking—Banish waste and create wealth in your corporation*. London: Simon & Schuster.

Womack, J.P., Jones, D.T., & Roos, D. (1990). *The machine that changed the world*. New York: Harper Perennial.

Wysocki, B. (2004). Industrial strength: To fix health care, hospitals take tips from factory floor; adopting Toyota techniques can cut costs, wait times; ferreting out an infection; what Paul O' Neil has been up to. *The Wall Street Journal*, April 9.

KEY TERMS

Artificial Variability: The poor management of the processes used to provide care. This type of variability (dysfunctional management and policies) can be reduced.

Lean Principles: Lean principles were originally developed in Toyota's manufacturing operations, known as the Toyota Production System. The term was popularized in the seminal book, *The Machine that Changed the World* (Womack et al., 1990), which clearly illustrated for the first time the significant performance gap between the Japanese and Western automotive industries.

Natural Variability: A source of great waste in the health care delivery system is excessive variability in the processes used to provide care. Natural variability is largely outside the control of a hospital. It includes clinical variability (patients differ in the type and severity of their diseases, and similar patients respond differently to treatment), patient demand variability (patients arrive for treatment randomly over time), and professional variability (different providers treat similar patients in different ways), which has given rise to the development of approaches like practice guidelines and clinical pathways.

Patients in Progress (PIP): In a hospital environment, patients in progress (PIP) represent the number of patients in the system at one time. PIP is a form of inventory, and the level of PIP is determined by batching considerations and variability.

Patient Resident Time: The amount of time a patient spends in the health care system, typically measured from the moment he or she enters the system until time of medical discharge.

Wasteful Work: Wasteful work is the fraction of the total time and effort in any organization that *does not* add value for the end customer. By clearly defining value for a specific product or service from the end customer's perspective, all the nonvalue activities, or waste, can be targeted for removal step by step. For most production operations, only 5% of activities add value; 35% are necessary nonvalue-adding activities, and 60% add no value at all. Eliminating this waste is the greatest potential source of improvement in corporate performance and customer service.

Examples of waste:

- Overproduction
- Waiting
- Transporting
- Inappropriate processing
- Unnecessary inventory
- Unnecessary movement

Clinical Decision Making by Emergency Room Physicians and Residents

Ken J. Farion
Children's Hospital of Eastern Ontario, Canada

Michael J. Hine
Carleton University, Canada

Wojtek Michalowski
University of Ottawa, Canada

Szymon Wilk
Poznan University of Technology, Poland

INTRODUCTION

Clinical decision-making is a complex process that is reliant on accurate and timely information. Clinicians are dependent (or should be dependent) on massive amounts of information and knowledge to make decisions that are in the best interest of the patient. Increasingly, information technology (IT) solutions are being used as a knowledge transfer mechanism to ensure that clinicians have access to appropriate knowledge sources to support and facilitate medical decision-making. One particular class of IT in which the medical community is showing increased interest is clinical decision support systems (CDSSs).

CDSS is "any program designed to help healthcare professionals make clinical decisions" (Musen, Shahar & Shortliffe, 2001). Decision models used in CDSS, especially those providing patient management and diagnostic advice, are normally based on expert knowledge, either discovered from past data or elicited from medical books or practice guidelines. The quality of any patient-specific CDSS is reliant on the quality of the underlying decision model(s). These models have to reflect clinical expertise, which implies that clinicians using such systems have to provide values for the CDSS input variables that can be correctly elicited only with an appropriate level of expertise. That is, only experienced clinicians will be able to provide CDSS input variables in a reliable and comprehensive manner, while inexperienced clinicians will be forced to gather information and make assessments for activities that they may lack the clinical acumen to do accurately. This may diminish the usefulness of the CDSS and the validity of the advice generated by the system, and lead to the rejection of the system by novice clinicians as forcing them to evaluate a patient in a way in which they are not accustomed.

CDSS users can be categorized using the classical taxonomy of novice or expert decision-makers. Differences between these two classes of decision-makers have been widely documented in the decision-making and medical literature. In complex domains such as medicine, it typically takes 10 years of training before one can be considered an expert (Prietula & Simon, 1989). Over time, experts develop a capability to systematize information and to form complex networks of knowledge that is stored in long-term memory (Arocha, Wang & Patel, 2005; Prietula & Simon, 1989). Novices lack these knowledge networks, and thus, when faced with new informational cues, they need to produce more hypotheses than experts (Kushniruk, 2001) and are unable to filter out irrelevant cues (Patel, Arocha & Kaufman, 1994), thus taking a longer time to make their decisions.

The purpose of this chapter is to explore how two classes of CDSS users representing different levels of expertise consider expert-generated CDSS inputs in their clinical decision-making. In this study, staff physicians are considered expert decision-makers and residents are considered novice decision-makers. Our study is based on the empirical results of a clinical trial of a CDSS that was developed for helping with triage decisions of pediatric abdominal pain in an emergency department (ED) (Farion, Michalowski, Slowinski, Wilk & Rubin, 2004). On the basis of collected data and other established literature on expert/novice decision-making, we evaluate differences between these two groups of CDSS users and draw more general conclusions for supporting clinical decision-making with technology.

The research question we seek to answer is:

What importance do residents and staff physicians place on expert-generated CDSS input variables in making their clinical decisions?

This chapter is organized as follows. First, background about the Mobile Emergency Triage (MET) CDSS is presented, along with an explanation of the input variables that are used by the system. Next, descriptions of the sample and data collection procedures are provided, along with the analysis techniques used. This is followed by a discussion of the results, future trends in CDSS design, and a conclusion.

BACKGROUND

The MET system was designed and developed to support ED physicians in making triage decisions about children with abdominal pain. The MET system consists of a server that interfaces with a hospital's electronic patient record system using the HL7 protocol and a client that resides on a PDA. The client facilitates the collection of clinical data during examination by physicians and also supplies a triage support function. The client is used directly at the point of care.

The MET client provides a series of interfaces to collect 11 out of 13 input variables shown in Table 1 that are used by the abdominal pain triaging algorithm (the remaining two variables, gender and age, are extracted automatically from the electronic patient record system). The collected data get transferred to the server for persistent storage and usage in the electronic patient record system. Discretizations for numerical input variables were based on medical practice, and the triage decision-making model was developed using retrospective chart analysis and knowledge discovery techniques based on rough set theory (Pawlak, 1991; Slowinski, 1995). The decision model is represented as decision rules that are easy to comprehend and interpret by physicians, and therefore are well accepted in clinical practice.

Based on the values of the input variables, the client uses the rule-based decision model to offer a suggested triage decision, which can be one of the following three options:

Table 1. Abdominal pain triaging input variables

Input Variable Name and Description	Possible Values
Age	number, discretized to 0-6, >= 7 years
Gender	male, female
Duration of pain	number, discretized to <=24 hrs, 1-7 and >7 days
Site of maximal pain	right lower quadrant (RLQ), lower abdomen, other
Type of maximal pain	continuous, other
Vomiting	yes, no
Previous visits in the ED for abdominal pain during the last 48 hours (irrespective of site)	yes, no
Temperature	number, discretized to <37, 37-39, >= 39 Celsius
Site of maximal tenderness	RLQ, lower abdomen, other
Localized guarding: localized muscle sustained contraction noted when palpating the abdomen	absent, present
Rebound tenderness: pain felt at site of maximal tenderness, produced by altering intra-abdominal pressure	absent, present
Shifting of pain	absent, present
WBC (white blood cells)	number, discretized to <=4000, 4000-12000, >=12000

- **Discharge.** Patient can be discharged to home, as their pain is caused by a nonserious problem.
- **Observation/Investigation.** Further in-hospital evaluation is required to determine the cause of the pain, as a serious cause (other than appendicitis) is likely.
- **Consult.** Surgeon is called because acute appendicitis is suspected.

Abdomen location input variables (site of maximal pain, site of maximal tenderness) were collected by clicking on an abdomen pictogram on the PDA. A screen capture showing the interface for "site of pain" is provided in Figure 1. The resulting input variable values (RLQ, lower abdomen, other) and corresponding areas on the abdomen were defined by surgeons and ED physicians. Figures 2 and 3 show MET screen captures for "type of pain" and "temperature," respectively.

METHODS AND RESULTS

This study of staff physician and resident decision-making was part of a larger clinical trial that was designed to evaluate MET-AP clinical accuracy with physicians' triage predictions. Results of that study can be found in Farion, et al. (2007).

Sample and Data Collection

A convenience sample of 574 eligible children, aged 1 to 16 years, with acute AP were enrolled with consent between July 2, 2003, and February 29, 2004, at the Children's Hospital of Eastern Ontario (CHEO) ED.

Figure 1. MET-AP screen capture for site of pain

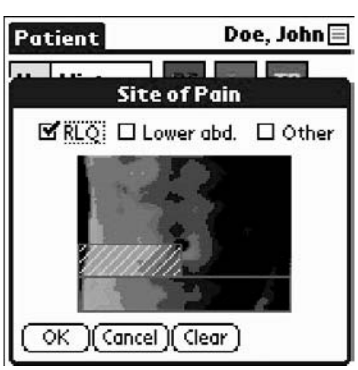


Figure 2. MET-AP screen capture for type of pain

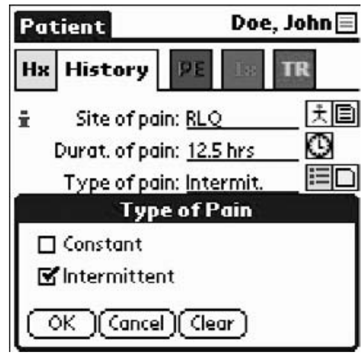
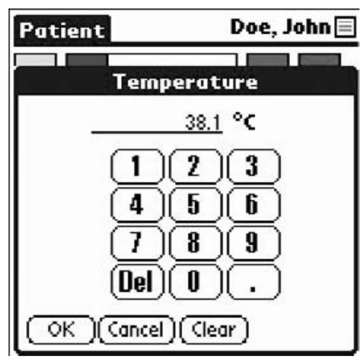


Figure 3. MET-AP screen capture for temperature



The treating ED resident or staff physician recorded their findings using MET-AP's electronic structured data screens. Blinded to the CDSS recommendation, the clinician entered his or her prediction of which triage category into which the patient was most likely to fit. This prediction was made at the time of initial assessment prior to obtaining an abdominal ultrasound or surgical consult, if required. A clinician of the opposite level (i.e., resident, staff physician) completed an independent interrater assessment within one hour of the original assessment, when possible.

Forty staff physicians and 110 residents enrolled patients. The ED clinicians had varying degrees of experience with handheld computers before entering the trial. All clinicians received in-depth orientation and training sessions, and as a result, all could use the system easily before the trial began. Two hundred and twenty two of the patients were seen by both a resident and a staff physician.

Analysis

The analysis focused on determining which of the CDSS input variables were being used in the clinicians' triage decisions. Since our independent variable (triage decision) is categorical, logistic regression was used to determine significant decision-making input variables. While triage decision initially had three categories, we have collapsed the observation and discharge result into a single category. We have done this because we are primarily interested in the input variables that are used to arrive at a consult decision, as this represents the most important clinical decision of the three possibilities.

Full main effects models were run independently for patients who were seen by residents and patients who were seen by staff physicians. All input variables except white blood cell count were included in the model. White blood cell count was removed from the study because of extensive missing data.

Results

The logistic regression results for residents and staff physicians are shown in Tables 2 and 3, respectively. The Nagelkerke's R^2 are .568 and .699 for the residents' and staff physicians' regression models, respectively. Overall, staff physicians have a higher number of significant CDSS input variables in their triage decision models than do residents. Specifically, pain site, pain type, vomiting, localized guarding, and rebound tenderness are significant for staff physicians. Alternatively, residents only had tenderness site, localized guarding, and rebound tenderness as significant variables. In terms of the "number of significant variables," these results are consistent with literature on strategic experts, which

Table 2. Logistic regression analysis for residents (n = 294 patients)

Variable	β	Std. Error	Wald Statistic	p-value
Age	0.498	0.994	0.251	0.617
Gender	-0.939	0.528	3.159	0.076
Pain Duration			0.325	0.850
Pain Duration (1)	-0.288	0.509	0.319	0.572
Pain Duration (2)	-5.306	63.417	0.007	0.933
Pain Site			0.153	0.926
Pain Site(1)	0.177	0.906	0.038	0.845
Pain Site(2)	0.440	1.124	0.153	0.696
Pain Type	0.692	0.511	1.833	0.176
Vomiting	0.035	0.487	0.005	0.944
Previous Visit	-6.895	29.973	0.053	0.818
Temperature			1.327	0.515
Temperature(1)	0.040	0.489	0.007	0.935
Temperature(2)	-1.911	1.695	1.271	0.260
Tenderness Site			9.971	0.007**
Tenderness Site(1)	2.741	0.944	8.427	0.004**
Tenderness Site(2)	0.361	1.305	0.076	0.782
Localized Guarding	1.863	0.508	13.469	0.000***
Rebound Tenderness	1.503	0.526	8.164	0.004**
Pain Shifting	0.766	0.514	2.222	0.136
Constant	-5.142	1.130	20.686	0.000
Nagelkerke R^2	0.568			

*Note: *p < 0.05, **p < 0.01, ***p < 0.001* states that experts have complex structures that assist in the recognition and interpretation of environmental signals and events (Lyles & Schwenk, 1992) and that these structures are more complex, contain more links among elements, and hence contain more elements then the cognitive structures of less experienced strategists (Day & Lord, 1992; Lurigio & Carrol, 1985).

Of note is the significance of input variables that are "most dependent" on physical examination (localized guarding and rebound tenderness) in residents' and staff physicians' regression models. This result is not surprising, given the educational focus of these input variables as being primary determinants of acute appendicitis. However, research has shown that residents often have deficiencies in their physical examination skills (Mangione, Burdick & Peitzman, 1995) and that novice physicians have generally weaker information gathering and decision-making skills than more experienced physicians (Johnson & Carpenter, 1986). We know from past empirical studies that clinicians with different levels of expertise exhibit differences in their abilities to collect and interpret information from physical examinations (Pines, Uscher Pines, Hall, Hunter, Srinivasan & Ghaemmaghami, 2005; Yen, Karpas, Pinkerton and Gorelick, 2005). In comparing abdominal examinations of ED pediatric patients undertaken by residents and attending physicians, it was found that all parts of the examination had less than moderate agreement (Yen et al., 2005). Similar results were found in studying abdominal examinations of nonminors by residents and attending physicians (Pines et al., 2005). While the "correctness" of eliciting the input variables was not explicitly validated in this study, the implication is that residents rely on input variables that they may be deficient at eliciting, in making their clinical decisions.

Table 3. Logistic regression analysis for staff physicians (n = 385 patients)

Variable	β	std. Error	Wald Statistic	p-value
Age	1.315	1.306	1.013	0.314
Gender	-0.593	0.528	1.260	0.262
Pain Duration			0.614	0.736
Pain Duration(1)	0.377	0.514	0.537	0.464
Pain Duration(2)	-5.517	20.305	0.074	0.786
Pain Site			6.862	0.032*
Pain Site(1)	2.467	0.973	6.429	0.011*
Pain Site(2)	2.376	1.381	2.960	0.085
Pain Type	1.611	0.614	6.879	0.009**
Vomiting	1.299	0.601	4.674	0.031*
Previous Visit	2.691	1.417	3.604	0.058
Temperature			2.312	0.315
Temperature(1)	0.619	0.534	1.343	0.246
Temperature(2)	2.421	2.097	1.333	0.248
Tenderness Site			3.194	0.203
Tenderness Site(1)	1.082	0.953	1.288	0.256
Tenderness Site(2)	-1.256	1.384	0.823	0.364
Localized Guarding	1.539	0.556	7.662	0.006**
Rebound Tenderness	2.306	0.576	16.005	0.000***
Pain Shifting	0.968	0.560	2.985	0.084
Constant	-8.380	1.692	24.533	0.000
Nagelkerke R^2	0.699			

Note: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$

FUTURE TRENDS

In evaluating the use of a CDSS for abdominal pain ED triage, we found that staff physicians used more of the CDSS input variables in their triage decision than did residents. The importance of the input variables that required physical examination was underlined by their significance in both staff physicians' and residents' regression models, even though past research suggests that residents have trouble accurately eliciting variables dependent on physical examination.

In order to take into account differences in clinical experience and to ensure appropriate support is available to these various user groups, we propose that the CDSS designers should (a) differentiate between information values provided by the data coming from expert and novice assessments, and (b) implement logical variable patterns that warn users when a single variable or a combination of variables is out of the expected range.

To design and implement aids that consider information value of the inputs, the input variables used in CDSS models must be categorized. Required variables could be logically categorized based on how difficult they are to elicit; to what extent they are reliant on tacit, explicit, and declarative knowledge; and subsequently be possibly labeled as "low confidence" and "high confidence" variables. This broad categorization reflects the ability of various physician user groups to accurately elicit values for the different attributes. A typical novice physician would have elicitation difficulty with "low confidence" attributes. Therefore, the user interface for the "low confidence" attributes should provide extensive explanations and guidelines to assist the process of collection. Further, provision for recording imprecise or uncertain information (e.g., selecting several values instead of a single one, entering some "confidence factor" associated with a value, or having a discrete option for "uncertain") should be provided.

In clinical decision-making, values of selected attributes form a pattern that is indicative of an underlying health condition. For example, for pediatric abdominal pain, certain pain location in concert with presence of guarding is indicative of possible acute appendicitis. It is possible to use information about such patterns to develop context-sensitive thresholds for values of individual attributes and their combinations. If values entered by a physician would significantly deviate from these logical thresholds, a CDSS would issue a specific warning alerting the physician to this situation. While this will provide additional support for novice physicians, it will also help to minimize the potential error between user and technology, which recently has been identified as an important source of clinical error (Kohn, Corrigan & Donaldson, 2000).

CONCLUSION

Many decision models implemented into CDSS encapsulate knowledge that relies on evaluating input variables that require experience and significant clinical acumen. This creates uncertainty about the quality of the recommendations produced by the CDSS. Customized decision support, taking into account the level of clinical expertise of a physician of a given specialty, is required to ensure that inputs into CDSS are accurate. Such expanded support is as important for the acceptance of a CDSS by physicians as the quality of the underlying decision model and user interface.

ACKNOWLEDGMENT

Research described in this chapter was supported by grants from NSERC-CHRP and Physician Services Inc. Foundation. A longer variation of this chapter was presented at the 12th Americas Conference on Information Systems; a longer version exists in the Conference Proceedings.

REFERENCES

Arocha, J.F., Wang, D., & Patel, V. (2005). Identifying reasoning strategies in medical decision-making: A methodological guide. *Journal of Biomedical Informatics*, *38*(2), 154–171.

Day, D.V., & Lord, A.G. (1992). Expertise and problem categorization: The role of expert processing in organizational sense-making. *Journal of Management Studies*, *29*(1), 35–47.

Farion, K.J., Michalowski, W., Rubin, S., Wilk, S., Correll, R., & Gaboury, I. (2007). Prospective evaluation of the MET-AP system providing triage plans for acute pediatric abdominal pain. *International Journal of Medical Informatics*, *77*(3), 208-218.

Farion, K.J., Michalowski, W., Slowinski, R., Wilk, S., & Rubin, S. (2004). Rough set methodology in clinical practice: Controlled hospital trial of the MET system. In S. Tsumoto, R. Slowinski, J. Komorowski, & J. Grzymala-Busse (Eds.), *Rough sets and current trends in computing* (pp. 805–814). Heidelberg: Springer Verlag.

Johnson, J.E., & Carpenter, J.L. (1986). Medical house staff performance in physical examination. *Archives of Internal Medicine, 146*(5), 937–941.

Kohn, L.T., Corrigan, J.M., & Donaldson, M.S. (2000). *To err is human: Building a safer health system*. Washington, DC: Institute of Medicine. National Academy Press.

Kushniruk, A.W. (2001). Analysis of complex decision-making processes in health care: Cognitive approaches to health informatics. *Journal of Biomedical Informatics, 34*(5), 365–376.

Lurigio, A.J., & Carrol, J.S. (1985). Probation officers' schemata of offenders: Content, development and impact of treatment decisions. *Journal of Personality and Social Psychology, 48*(5), 1112–1126.

Lyles, M.A., & Schwenk, C.R. (1992). Top management, strategy and organizational knowledge structures. *Journal of Management Studies, 29*(2), 155–174.

Mangione, M., Burdick, W.P., & Peitzman, S.J. (1995). Physical diagnosis skills of physicians in training: A focused assessment. *Academic Emergency Medicine, 2*, 622–629.

Musen, M.A., Shahar, Y., & Shortliffe, E.H. (2001). Clinical decision support systems. In E.H. Shortliffe, L.E. Perreault, G. Wiederhold, & L.M. Fagan (Eds.), *Medical informatics. Computer applications in health care and biomedicine* (2nd ed.) (pp. 573–609). New York: Springer.

Patel, V.L., Arocha, J.F., & Kaufman, D.R. (1994). Diagnostic reasoning and medical expertise. *Psychology of Learning Motivation, 31*, 187–252.

Pawlak, Z. (1991). *Rough sets: Theoretical aspects of reasoning about data*. Norwell, MA: Kluwer Academic Publishers.

Pines, J., Uscher Pines, L., Hall, A., Hunter, J., Srinivasan, R., & Ghaemmaghami, C. (2005). The interrater variation of ED abdominal examination findings in patients with acute abdominal pain. *American Journal of Emergency Medicine, 23*, 483–487.

Prietula, M.J., & Simon, H. (1989). The experts in your midst. *Harvard Business Review, 67*(1), 120–124.

Slowinski, R. (1995). Rough sets approach to decision analysis. *AI Expert Magazine, 10*, 8–25.

Yen, K., Karpas, A., Pinkerton, J.J., & Gorelick, M.H. (2005). Interexaminer reliability in physical examination of pediatric patients with abdominal pain. *Archives of Pediatrics & Adolescent Medicine, 159*, 373–376.

KEY TERMS

Clinical Decision Support Systems (CDSS): Software that assists clinicians in making health care decisions.

Emergency Department (ED): The department responsible for providing immediate medical care to patients arriving at the hospital.

Logistic Regression: Technique for making predictions when a dependent variable is a categorical dichotomy, and the independent variable(s) are continuous and/or categorical.

Pediatric Abdominal Pain Triage: A process of sorting children with abdominal pain into categories that reflect the treatment they require

Personal Digital Assistant (PDA): A handheld computer device.

A Collaborative Approach for Online Dementia Care Training

Colla J. MacDonald
University of Ottawa, Canada

Emma J. Stodel
Learning 4 Excellence, Canada

Lynn Casimiro
University of Ottawa, Canada

Lynda Weaver
SCO Health Service, Canada

INTRODUCTION

There are several reasons why university-based researchers and community groups may choose to collaborate together on research projects. Involving the end-users of the research data in the research process often motivates them to integrate the results into new policies, procedures, and education programs. Research outcomes therefore become more relevant to the community members than would be the case using a more traditionalistic approach to research (Morrison & Lilford, 2001; Patton 1997). Community-based partners fully immerse themselves in a collaborative research process as they strive to underpin their interventions with other complementary concepts or evidence-based theories. They are then better positioned to promote social change.

Furthermore, collaborative projects provide an educational opportunity for partners to develop a collective consciousness in addressing the issues at hand (Gallagher, Easterling, & Lodwick, 2003; Karim, 2001; Minkler & Hancock, 2003). The problem is viewed from multiple perspectives as university-based researchers and community professionals contribute unique strengths and share research-related responsibilities within the social and cultural dynamics of the partnership (Gibbon, 2002). Simultaneously, university-based researchers are able to come to a better understanding of the community of interest and its changing realities. The cultural differences of both groups are acknowledged, and sensitive strategies can be collaboratively developed in which the roles and expectations are clearly outlined (Agency for Healthcare Research and Quality, 2003). Full advantage can be taken of the knowledge, experiences, and perspectives of professionals in the community as they provide input on all aspects of the research project. Thus, the research process becomes a collaborative, co-learning, community-building experience.

There are obvious benefits to working in collaboration. However, real collaboration takes time; time to engage in meetings, complete accountability processes, and resolve problems. The delicate balance between democracy and efficiency can be compromised when you have to choose between equal participation and looming deadlines (Stoecker, 2003). Weaver and Cousins (2004) described this dilemma as assessing manageability or having to make a choice between achieving complete diversity on the researcher-community team and the unwieldiness of working with a large committee. Compromise is often necessary.

This article describes our experiences using a collaborative approach involving university-based researchers and community professionals—in this case, long-term care (LTC) managers, administrators, and hospital-based educators and researchers—to create an online dementia care training program.

BACKGROUND

The Sisters of Charity of Ottawa Health Service (SCOHS) is a corporation with a teaching chronic care hospital and two LTC facilities. Community professionals at SCOHS recognized that their healthcare providers were facing challenging behaviours from persons suffering from dementia, which has been known to lead to staff burnout, distress, and high turnover rates. The community professionals felt that this problem could be partially ameliorated with staff education through e-learning. They contacted a professor in the Faculty of Education at the University of Ottawa who was conducting research on e-learning. Together, they agreed that conducting a project involving frontline workers in LTC facilities that addressed this issue would be mutually beneficial. Additional experts were recruited to join the project, including psychologists as content experts, e-learning course developers and pedagogy experts, and evaluators experienced in online course evaluation. Six pilot LTC facilities were identified, and representatives from each facility were included in the research team.

THE ONLINE DEMENTIA CARE TRAINING PROJECT

The research group used the Demand-Driven Learning Model (DDLM) (MacDonald, Stodel, Farres, Breithaupt, & Gabriel, 2001, see Figure 1) to guide the design, development, delivery, and evaluation of the bilingual dementia care training program. The program was targeted towards frontline healthcare providers (registered and nonregistered) who care for persons experiencing dementia in LTC facilities.

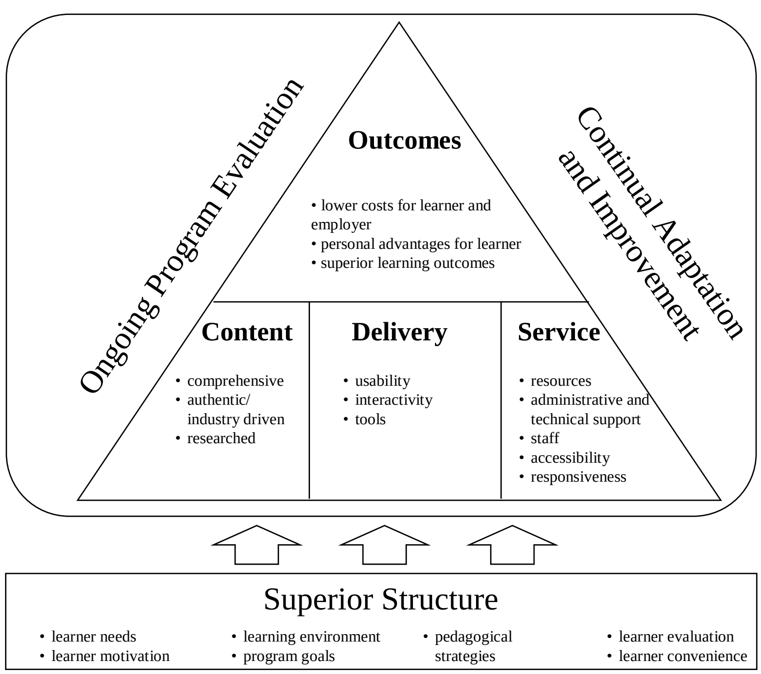
Conducting the Needs Assessment

As advocated in the DDLM, the first step of the project involved identifying the needs of the learners. The university-based researchers conducted three in-depth focus group interviews with seven healthcare providers (prospective learners) and two site coordinators who would serve as the on-site support persons for the learners during the implementation of the program. Through the needs analysis process, community-based stakeholders were able to provide input regarding the design, development, and delivery of the dementia care training program. The results of the needs analysis are published elsewhere (MacDonald, Stodel, & Coulson, 2004).

Designing and Developing the Program

The course facilitator, a content expert and psychogeriatric nurse with no e-learning experience, was responsible for putting together the first draft of the program content and learning activities. To ensure the

Figure 1. The demand-driven learning model (MacDonald, Stodel, Farres, Breithaupt, & Gabriel, 2001)



content was authentic to the community experience, as recommended by the DDLM, the course facilitator drew on her experiences working with frontline personnel who provide care to individuals with dementia. Once she developed the content, the pedagogy team reworked it into an appropriate format for e-learning. This was a time-consuming process, but one the academics felt strongly about as they were concerned with respecting online pedagogical principles and the quality of the online course design.

The pedagogy team encountered competing tensions as they developed the content using a collaborative approach. Ensuring the content was comprehensive, another DDLM component, was difficult to achieve due to time limitations. The content had to be chunked into sections that could be completed in 30-minute learning sessions, which limited the amount of content that could be included. Thus, content experts were advocating for the inclusion of a wide range of content, while other stakeholders wanted to reduce the amount of content, as they were aware of the limited time the learners would have to engage in the program.

Other issues related to time also affected the collaborative process. The pedagogy team had limited time to develop the content and adapt it to an online format, due to deadlines imposed by the funding agency and delays in hiring qualified project staff. Once the content was approved, it was given to the instructional designer to be put online. The instructional designer developed the online program using WebCT, the online course management system adopted by the University of Ottawa. However, the pedagogy team quickly realized that, while having many advantages, this system limited certain aspects of course design and evaluation. Time constraints and the initial decision to work with the University of Ottawa's Centre for E-Learning prevented the research team from exploring alternatives to WebCT.

The instructional designer questioned the efficiency of the collaborative process adopted in this project. He felt there were too many people trying to reach consensus. He postulated the lack of efficiency was because "the roles of the people involved were not well-defined." Similarly, not all stakeholders felt they were heard when they contributed expertise. One team member stressed that learners must be honestly informed about the time commitment required to participate in this project, prior to their consent and participation. Another found the program hard to navigate, and suggested changes be made in the online layout. However, despite these warnings, neither the amount of content was reduced to meet the learners' expectations and needs, nor the navigation improved before the program was offered, and both were identified as weaknesses of the program in the evaluation (MacDonald, Stodel, & Casimiro, 2006). These changes did not occur due to time constraints and/or lack of project ownership. Compromise, rather than consensus, became the decision method of choice, especially as the deadlines loomed closer.

A significant factor in the lack of project ownership was that the principal investigator had to leave the project suddenly at the beginning of the production phase. This temporarily compromised the project's leadership. Leadership is necessary for the smooth running of a project. This leader should bring expertise in optimizing group functioning, such as defining roles and orienting the group towards action. This person should also ensure equality among members and aim for consensus.

Delivering the Program

The completed program was delivered to 95 learners at six sites in three provinces across Canada. Forty-nine (52%) enrolled in the French language program, and 46 (48%) enrolled in the English language program. Learners were expected to spend two hours each week, at their convenience, reading the content and completing the learning activities and evaluations. Each site had a coordinator whose role was to support the learners in their learning and with the technology. These site coordinators were an integral part of the pedagogy team.

Throughout the delivery of the program, the site coordinators met with the pedagogy team for 30 minutes each week via teleconference. These meetings enabled site coordinators to keep track of the timelines set out for the learners, and allowed them to provide feedback to the pedagogy team regarding the learners' experiences and progress with the program. These meetings were instrumental to the success of the program. Because of the regular feedback from the site coordinators, the pedagogy team was able to address problems and concerns that arose in an expedient fashion. For example, when it became apparent that the program required significantly more than two hours a week to complete, and that some of the learners felt frustrated and overwhelmed, the pedagogy team promptly adapted the program by reducing the number of required exercises and extending the deadlines to complete the program. A number of site coordinators attested that the learners immediately felt less pressure once the amount of work was reduced.

Evaluating the Program

One of the reasons the community created a partnership with academics was to help develop strategies for evaluating the program. The university-based researchers had specific knowledge of evaluation and online data collection procedures. Using the DDLM evaluation tool (MacDonald, Breithaupt, Stodel, Farres, & Gabriel, 2002) as a guide, they drafted the evaluation instruments and invited the pedagogy team to provide input. As a result, survey questions were added, deleted, or modified. This collaborative process resulted in the development of relevant evaluation tools tailored to the needs of learners in LTC facilities. Further, the evaluation team conducted in-depth semistructured interviews in order to obtain a rich description of the stakeholders' experiences with the program, specifically in terms of the DDLM components. Interviews were conducted with ten learners, all six site coordinators, the course facilitator, and the instructional designer. Knowing the importance of management buy-in for this program to be viable in the future, the pedagogy team also interviewed the higher management of the six LTC facilities to gather their impressions of the value of the program for their organizations.

The evaluation team conducted, transcribed, and analyzed all the interviews, compiled and analyzed both the quantitative and qualitative data (French and English), and wrote the final report (MacDonald & Stodel, 2004). The evaluation continued to be a collaborative process as both community professionals and university-based researchers were part of the evaluation team. Moreover, the evaluation team received input from the pedagogy team. This decision making process worked well; the group weighed the importance of outcomes relevant for the learners' practice—a reality best understood by the community—against the outcomes related to the e-learning experience—a reality best understood by the university-based researchers.

Disseminating Findings

The university-based researchers took the lead role in the dissemination activities. However, input received from community members was critical to ensure the quality and relevance of the final products. Input from community members allowed the university-based researchers to create a PowerPoint presentation and two professional posters comprising an overview of the project and its findings that could be used by the community to share details of the project at conferences or in their local healthcare communities. While these individuals may not have had the time or the interest to prepare the presentation material, they were able to actively participate in the dissemination process by using materials created by the university-based researchers. This highlights another benefit of taking a collaborative approach.

Writing this article involved a collaborative effort between three university-based researchers and one community member. The collaborative approach added to the amount of time necessary to complete the manuscript, yet it added to the quality and accuracy of the article on several fronts. First, both the university- and community-based members' perspectives were reflected. Second, perceptions of the process of running the project were shared and elaborated. Lastly, the workload was shared between the authors. In the end, the process of collaborating in the publication process was a positive experience in which co-learning took place, and a stronger article resulted.

FUTURE TRENDS

The implications and lessons learned from this research are important to consider when conducting future research in this area. Although several of the community members were experienced at and capable of conducting research and disseminating the findings without the assistance of the university-based researchers, they felt a partnership with academics would strengthen their research position in the field of e-learning, and they would gain information on quality e-learning design. In turn, the academics wished to apply their knowledge to the healthcare field, but lacked the intimate knowledge of this group of learners. Consequently, they benefited from the input of the community members and gained valuable knowledge of the content area (dementia care) and healthcare culture. Even though the university-based researchers had much to gain from the partnership, this project was instigated and led by members of the community.

Our experience suggests that taking a collaborative research approach significantly increases the amount of time required to complete a project, though the research team felt it was a worthwhile trade-off, given the contributions made by each member. We acknowledge that failure, at times, to reach consensus and the pressure of timelines to complete this project by a specific date compromised the quality of the design. From an e-learning perspective, it is critical that the content be developed early in the process. A common fallacy is that once the content is developed, it can be put online quickly.

Stoecker (2003) suggested that it is an unrealistic expectation that those involved in a collaborative project be equal partners in all aspects of the process. His claim appears to be grounded in concerns for the time demands on community members. Indeed, as Israel, Schulz, Parker, Becker, Allen, and Guzman (2003) have suggested, efforts were made to involve community members in the publications resulting from the research process, but few had the time to do so. Sullivan, Chao, Allen, Koné, Pierre-Louie, and Krieger (2003) recommended that communities receive concrete benefits in return for their involvement in research partnerships, noting that without such tangible benefits, the partnership may not be advantageous to the community.

We also realized how important communication is to successful collaboration. Working in a large team means that many relationships need to be developed and maintained. In this project, constructive criticism about the program was sometimes taken as a personal affront by the individual responsible for that aspect of the program, and hard feelings resulted. This point becomes even more poignant when communication between members is conducted via e-mail where tone, emotion, and other nonverbal cues are lost. In the case of this project, misunderstandings in communication caused some minor conflicts, but effective communication also allowed the rifts to be resolved.

Clarification of members' roles at the beginning of the project is also critical. In writing this article, it emerged that the evaluation team was ill-defined; each of us had a different view of whom the evaluation team comprised. Further, one of the pedagogy team members who had experience in online pedagogy and e-learning applications in healthcare felt that her expertise was not used as much as it could have been. Not only was this a frustrating experience for her, but she also predicted many of the problems reported during program delivery.

In the end, the project resulted in an educational opportunity for all involved (Minkler & Hancock, 2003). The community of program personnel succeeded in offering a generally well received e-learning program to help healthcare providers manage persons with dementia. They also enhanced their skills for designing, developing, delivering, and evaluating a successful e-learning program by using the DDLM. The university-based researchers learned more about healthcare in general, and dementia care specifically, obtained further appreciation for healthcare providers' needs for learning online, and became aware of the values and complexities of collaboration. Further, the university-based researchers also profited by gaining access to data for research purposes.

CONCLUSION

In sum, the adoption of a collaborative research approach with a community of professionals is not only beneficial but also desirable. Collaboration allowed multiple views, attitudes, and experiences to strengthen the program. By describing the complexities involved in this process, we reveal both the challenges and achievements inherent in designing a quality online learning event through collaboration. Moreover, by addressing some of the details involved in this process, we hope that our experiences help others plan collaborative partnerships with community professionals, and develop collaborative e-learning programs for healthcare providers.

REFERENCES

Agency for Healthcare Research and Quality. (2003) *The role of community-based participatory research: Creating partnerships, improving health.* AHRQ Publication No. 03-0037. Rockville, MD: U.S. Department of Health and Human Services. Available online at http://www.ahrq.gov/research/cbprrole.pdf

Gallagher, K. M., Easterling, D. V., & Lodwick, D. G. (2003). Introduction. In M. Minkler, & N. Wallerstein (Eds.), *Community based participatory research for health* (pp. 1–16). San Francisco, CA: Jossey-Bass.

Gibbon, M. (2002). Doing a doctorate using a participatory action research framework in the context of community health. *Qualitative Health Research, 12*(4), 546–558.

Isreal, B. A., Schulz, A.J., Parker, E. A., Becker, A. B., Allen, A. J., & Guzman, J. R. (2003). Critical issues in developing and following community based participatory research principles. In M. Minkler, & N. Wallerstein (Eds.), *Community based participatory research for health* (pp. 53–76). San Francisco, CA: Jossey-Bass.

Karim, K. (2001). Assessing the strengths and weaknesses of action research. *Nursing Standard, 15*(26), 33–35.

MacDonald, C. J., Breithaupt, K., Stodel, E. J., Farres, L. G., & Gabriel, M. A. (2002). Evaluation of Web-based educational programs: A pilot study of the demand-driven learning model. *International Journal of Testing, 2*(1), 35–61.

MacDonald, C. J., & Stodel, E. J. (2004). *An e-learning dementia care program for healthcare workers in LTC facilities: Final evaluation report.* Unpublished manuscript.

MacDonald, C. J., Stodel, E. J., & Casimiro, L. (2006). Online dementia care training for healthcare teams in continuing and long-term care facilities: A viable solution for improving quality of care and quality of life for residents. *International Journal on E-Learning, 5*(3), 373–399.

MacDonald, C. J, Stodel, E. J., & Coulson, I. (2004). Planning an e-learning dementia care program for healthcare teams in long-term care facilities: The learners' perspectives. *Educational Gerontology: An International Journal, 30*(10), 1–20.

MacDonald, C. J., Stodel, E. J., Farres, L. G., Breithaupt, K., & Gabriel, M. A. (2001). The demand-driven learning model: A framework for Web-based learning. *The Internet and Higher Education, 4*(1), 9–30.

Minkler, M., & Hancock, T. (2003). Community-driven asset identification and issue selection. In M. Minkler, & N. Wallerstein (Eds.), *Community based participatory research for health* (pp. 135–179). San Francisco, CA: Jossey-Bass.

Morrison, B., & Lilford, R. (2001). How can action research apply to health services? *Qualitative Health Research, 11*(4), 436–449

Patton, M. Q. (1997). *Utilization focused evaluation* (3rd ed.). Thousand Oaks, CA: Sage.

Stoecker, R. (2003). Are academics irrelevant? Approaches and roles for scholars. In M. Minkler, & N. Wallerstein (Eds.), *Community based participatory research for health* (pp. 98–112). San Francisco, CA: Jossey-Bass.

Sullivan, M., Chao, S. S., Allen, C. A., Koné, A., Pierre-Louie, M., & Krieger, J. (2003). Community-researcher partnerships: Perspectives from the field. In M. Minkler, & N. Wallerstein (Eds.), *Community based participatory research for health* (pp. 113–130). San Francisco, CA: Jossey-Bass.

Weaver, L., & Cousins, J. B. (2004). Unpacking the participatory process. *Journal of Multidisciplinary Evaluation, 1*, 19–40.

KEY TERMS

Collaborative Research Approach: A research approach that uses the strengths of both researchers and community-based practitioners to increase the quality of the research process and the impact of the research outcomes.

Demand-Driven Learning Model: An e-learning model that is grounded within a constructivist framework and defined by five interrelated dimensions that, in concert, create a high-quality e-learning experience: superior structure; three consumer demands of content, delivery, and service; and learner outcomes.

E-Learning: Learning that takes place via the Internet.

Long-Term Care Facility: A licensed residence for individuals who require personal support and nursing care, and who cannot remain in their family home, because support at home (from family or agencies) is insufficient or unavailable. While funding sources may differ, LTC facilities and nursing homes can be categorized together under this definition.

Instructional Designer: Person who uses technology (media) to design optimised learning events. Instructional design is historically grounded in cognitive and behavioural psychology.

Pedagogy: The profession, art, science, theory, principles, or methods of education, instruction, and teaching.

Psychogeriatric Nurse: A nurse whose practice is focused on older people (usually 65 years of age or older) with mental illness and/or cognitive impairment.

Web Course Tools (WebCT): An e-learning platform and online course management system used extensively in colleges, universities, and other educational institutions. WebCT supports online tools such as discussion forums, e-mail, live chat, and whiteboarding, as well as content in various formats (e.g., html documents, Web pages, and so on). WebCT recently merged with Blackboard, another leading provider of educational software.

C

Combining Technology with Tradition to Effect Superior Pain Management Strategies

Choong Khean Foo
RMIT University, Australia

INTRODUCTION

"Pain is a very good servant but a bad master."
- Author unknown

Everybody knows what pain is, but no one really knows how to switch it off, especially when it becomes persistent and annoying.

Pain is not just from physical disorders but also from combinations of physiological, pathological, emotional, psychological, cognitive, environmental, and social factors. The keys to successful pain control are the mechanisms that initiate and maintain pain. … Now, the public and health professionals expect to control pain by using preventive and active strategies, including drugs and physical and psychosocial interventions. (Holdcroft & Power, 2003)

Have we lost sight of the real pathophysiology that underlies the phenomenon of pain? Are we concentrating on the end point rather than the cause of the malady? Drugs and physical therapies don't have all the answers to pain management. Is there another way? Maybe we should take a broader look at why the pain started in the first place, and then we can work out the solution to the health challenge.

BACKGROUND

Pain arises when a living tissue is injured, and the tissue responds with inflammation. Inflammation involves a cascade of events involving cells, enzymes, cytokines, chemokines, and other substances, which will initiate other secondary responses by other cells and blood vessels. Inflammation is the beginning of the process of repair and healing. Ancient civilizations used various modalities to manage pain. They used physical therapies, medicinal foods, and herbs where appropriate.

Over the past century, pain and its management have changed dramatically. In this chapter, I propose to discuss the reasons for the change in pain experience and will explore the practical aspects of pain management using the best from the East and West. While scientists are dwelling deeply on the complex issues of pain and inflammation, patients in pain need help and advice now. Drugs alone cannot relieve all pain and suffering; surgical and psychological treatment may help some, but for most, a holistic approach may help to reduce or relieve the pain and suffering.

MAIN FOCUS OF THE CHAPTER

"We are what we eat."

If we supply our body with inappropriate or inferior quality materials, the tissues and structures so made will be of inferior quality. Deficiency diseases like scurvy and rickets have been known to cause deformed tissues and cause pain with different characters. By correcting the deficiencies, these maladies can be cured. In our modern society, it seems that more people are experiencing and suffering from pain. Is it because the newer generations are made of "weaker" materials, or is it that the pain threshold is reduced or the background inflammation in the body system has heightened? The fact that we have not been able to quantify and identify the missing ingredients in chronic pain should not deter us from searching for an answer by going "back to basics." Our body is created to function efficiently and effectively when it is provided with all the nutrients it needs.

MODERN MEDICAL MODEL OF PAIN MANAGEMENT

In modern medicine, pain is treated with analgesics and nonsteroidal anti- inflammatory drugs. If pain persists, stronger narcotics and anti-inflammatory corticosteroids are used. Other drugs like antidepression and anti-epilepsy drugs often are used in pain clinics to help modulate the pain experience. Unfortunately, all these medications have unwanted side-effects. There is a growing awareness in the community that drug therapy is not the only way. People are looking for more natural alternatives and less harmful treatment modalities with more self-help involvement. They want to participate actively in the healing process. People are turning more and more to alternative and natural therapies such as physical therapies, (e.g., physiotherapy, osteopathy, chiropractic) with stretching, mobilization, manipulations, heat/cold packs, and acupuncture/acupuncture-related techniques, along with diet and nutritional supplements, to help manage their painful conditions.

TRADITIONAL MEDICAL MODEL OF PAIN MANAGEMENT

In traditional Chinese medicine (TCM), diseases develop when the flow of Qi (energy, prana, life force) is blocked or interrupted. The basis of TCM treatment is to reestablish the flow of Qi. The TCM practitioners use needle acupuncture and/or herbal remedies to alter or reestablish the flow of Qi to reestablish homeoestasis in the body (Chinese acupuncture and moxibustion, 1990; O'Conner & Bensky, 1983).

The Qi or Life Force is also derived from the air we breathe and the water and food we drink and eat. Proper breathing and nutrition form part of the total management in TCM to help bring the body back into balance. It is well known that the food we eat determines how we feel. There are certain foods that will initiate or aggravate the painful situation. Food sensitivities and intolerance can trigger inflammatory response and cause pain (Author's clinical observation and treatment protocol).

ALTERNATIVE THERAPIES

Acupuncture, LLLT, & TENS

Since 1974, acupuncture has been an accepted treatment method in pain clinics around the world. There is only one slight drawback with acupuncture. Not many people, given a choice, would like to have needles stuck into them. So scientists did intensive research for a noninvasive acupuncture-point stimulation methods. In the East, traditional health practitioners used manual massage—Shiatsu, AnMo, and Tui Na—to stimulate the tissues and acupoints. Scientists discovered two modalities that can stimulate acupoints through the skin without inflicting pain by using Low Level LASER Therapy (LLLT) (Oshiro & Calderhead, 1990; Wong & Fung, n.d.; Baxter, 1994) and Transcutaneous Electrical Nerve Stimulation (TENS). Both these techniques are well accepted by the patients, as needles are not used. There is evidence that these methods of acupoint stimulation may give some of the benefits of needle acupuncture treatment. Although TENS was introduced more than 30 years ago as an alternative therapy to pharmacological treatments for chronic pain and in spite of its widespread use, the effectiveness of TENS is still rather controversial. A meta-analysis was carried out by scientists in 2001, and the results were published in the *Cochrane Database Systemic Reviews*, which showed that there was no evidence to support the use of TENS in the treatment of chronic low back pain, while the results for chronic pain were inconclusive.

Convention TENS stimulation has its limitation because of habituation from repetitive monotonous stimulation over the same points (Godfrey, n.d.; Carroll et al., 2001; Milne, et al., 2001). However, the Canadian research scientists were able to overcome this habituation by using a novel machine called Codetron®, which uses six negative electrodes and one positive electrode. At any one time, a pair of electrodes is stimulated. The sequence of stimulation is randomized so the brain will not filter out the stimulation and will accept each individual stimulation as a novel one. Controlled studies were done comparing Codetron with Sham Codetron therapy (Fargas-Babjak, Rooney, & Geretz-Simon, n.d.) and Codetron with Electroacupuncture (Richard & Cheng). The findings were very favorable toward Codetron therapy.

DIET, NUTRITION, AND PAIN MANAGEMENT

In recent years, many reports have been published stating that certain types of food cause or aggravate arthritic or rheumatic pain. Childers and Margoles (1993) in their article "An Apparent Relation of Nightshades (Solanaceae) to Arthritis" suggested that Solanaceae can cause pain in at least two known ways. First, the alkaloids in these foods can act alone or with other *cholinesterase inhibitors* (caffeine or food impurities containing systemic cholinesterase inhibiting pesticides) and can cause muscle spasms, aches and pain, tenderness, inflammation, and stiffness of body movements. Second, Solanaceae can develop the very active metabolite of vitamin D3 (1a25 dihydroxycholecalciferol), which can cause calcinosis of soft tissues, ligaments, and tendons, mineralization in walls of major arteries and veins, and osteopetrosis and related pathology in livestock. Over a period of time, these cause progression to lameness and extended uselessness, with eventual death of livestock (Childers & Margoles, 1993).

It was reported that 28% of arthritis sufferers had a "marked positive response," and another 44% had a "positive response' when they eliminated nightshades from their diet over a period of a few months (Childers & Margoles, 1993). In my clinical practice, a high percentage of rheumatic and arthritic patients improved just by avoiding the nightshade plants in their diet. The response can be as quick as a few weeks (Author's clinical observation and treatment protocol).

Although rheumatoid arthritis (RA) is listed as an autoimmune disease, it may be linked to food allergies and sensitivities (Zeller, 1949). In some RA patients, the pain is made worse when they eat foods to which they are allergic or sensitive, and vice versa, when these foods are avoided, the patients feel better (Author's clinical observation and treatment protocol; Beri et al., 1988; Panush, 1988; Taylor, 1983; Darlington & Ramsey, 1991). It is estimated that one-third of RA patients can control the disease completely through allergy elimination (Darlington & Ramsey, 1991).

Inflammation and Foods

Inflammation, with the accompanying features of redness, swelling, warmth, and pain, is the basic way living tissues react to injury, irritation, or infection. This is a form of nonspecific immune response.

The foods we eat have a powerful bearing on our health and, specifically, inflammation. Some foods we eat are pro-inflammatory, while others are anti- inflammatory. With the advent of fast food industries, our normal balanced inflammatory index diet has shifted toward 30 times more pro-inflammatory when compared with our forefathers' diets. As a consequence, nowadays it is not surprising to note that more people are suffering from various rheumatic-type pain. One of the contributing factors could be the types of foods we eat and how we prepare our foods. To win the war against inflammation, we have to turn the tide and go back to eating what Mother Nature wants us to eat—fresh living natural foods!

Fats, Oils, and Anti-Inflammatory Foods

Extremely low-fat diets have been reported to help people with RA (Lucas & Power, 1981), and very low-fat, pure vegetarian diets have also proved helpful (Skoldstram, 1987; Nenonen, Helve & Hanninen, 1992). While polyunsaturated fats like n-3 fatty acids and n-6 fatty acids of evening primrose oil are helpful, oily fatty greasy fried foods are known to be pro-inflammatory and can aggravate inflammation and intensify pain. Vegetable oils heated to high temperature can turn into pro-inflammatory oil. There are many foods that are known to be pro-inflammatory in nature, and whenever possible, they should be avoided by people who suffer from aches and pains. Some of the common foods are milk products, wheat and wheat products, red meat, eggs, preservatives, additives and colorings, sugars, and caffeine drinks (Author's clinical observation and treatment protocol; Warmbrand, 1974; Kjeldsen-Kragh, 1991).

In a laboratory experiment, Tall & Raja demonstrated that rats fed with soy protein diet had significantly less inflammation in their paws following an injection of Freud's adjuvant as well as having a higher tolerance to heat compared with those rats fed with casein-fed animals. These results are consistent with previous research showing consumption of a soy-containing diet suppressed the development of pain following nerve injury (Shir, Raja, Weissman, Campbell & Ze'ev Seltzer, 2001).

Preliminary evidence suggests that consumption of olive oil rich in oleic acid may decrease the risk of developing RA (Linos et al., 1999). One trial in which people with RA received either fish oil or olive oil, found those olive oil capsules providing 6.8 g of oleic acid per day for 24 weeks produced modest clinical improvement and beneficial changes in immune function. However, as there was no placebo group in that trial, the possibility of a placebo effect cannot be ruled out (Kremer et al., 1990).

Our forefathers used to eat food containing a higher proportion of n-3 fatty acids. It was suggested that the ratio of n-3 fatty acids—n-6 fatty acids—was in the region of 1:1. Nowadays with intensive farming and chemical fertilization, the ratio of n-3 fatty acids—n-6 fatty acids—is closer to 1:30. Possibly because of the skewed ratio, more people are suffering more aches and pains in our society. To help correct this imbalance, it will be prudent for us to eat more n-3 fatty acids, which are found in deep-sea fish and flaxseed oil.

Omega 3 fatty acid is used by the cells to keep cellular and nuclear membranes soft and pliable in order for the cells to function properly. Omega 3 fatty acid is also converted into Prostaglandin E(3) (PGE 3), a potent anti-inflammatory eicosanoid. Omega-6 fatty acid, which is found in vegetable oil, is converted to either Prostaglandin E (1) (PGE1) or Prostaglandin E (2) (PGE2). PGE1 is a strong anti-inflammatory eicosanoids, while PGE (2) is pro-inflammatory eicosanoids. Through proper diet and supplements of flaxseed oil and fish oil, it is possible to improve the n-3:n-6 fatty acids ratio and increase anti-inflammatory eicosanoids in the body (Author's clinical observation and treatment protocol; James, Gibson & Cleland, 2000).

Free Radicals and Antioxidants

Many painful conditions such as arthritis, fibrositis, and fibromyalgia are oxidative disorders due to free radical attacks. Increased free radical levels may be responsible for the development of many painful inflammatory conditions such as rheumatoid arthritis, fibromyalgia, osteoarthritis, and various rheumatic painful conditions (Packer, Hiramatsu & Yoshikawa, 1999; Bagis et al., 2005).

There is a wide range of antioxidants found naturally in the foods we eat and the herbs and spices we use for cooking. The most well known ones are Vitamin C and Vitamin E complex, which play a major role in supporting the other antioxidant activity in the body. Free radical attack causes damage to the cell structure by disrupting the stable cellular membranes. Inherently, a free radical is an unstable molecule with one or two missing electrons. Free radicals will try to snatch an electron from another stable molecule, and by so doing will cause a chain reaction of free radical attacks on the tissues, leading to damage to the tissue structure. The body tissues have an endogenous antioxidant system and will try to repair the free radical tissue damage. When DNA is damaged by the free radicals and it cannot be repaired, the disrupted genetic coding will duplicate without any central control. This is how cancer cells develop and multiply.

Supplements with vitamins, minerals, and phytonutrients such as carotenoids, flavonoids, and cruciferous, which contain potent antioxidants, have proved to be very effective in quenching the free radical attacks. When we eat a wide variety of fresh colorful fruits and vegetables, we are providing our body with a basket of antioxidant that will help neutralize the free radical activities and protect the cells from further damage (Packer, Hiramatsu & Yoshikawa, 1999).

In recent years, scientists have been looking at various phytonutrients in foods and plants that we consume in our diet. One such product is Pygnogenol (Procyanidins and Catechin) extracted from the French maritime pine bark. It is a very potent antioxidant and is also a strong anti-inflammatory agent, which will help to relieve many painful conditions. Anthocyanins, procyanidins, and quercetins are also found in grape skins and grape seeds. Another antioxidant that has proved to have both anti-oxidant and anti-inflammatory properties and has given great relief to people with various pain conditions is Xanthones. Xanthones are flavonoids and are found in some fruits and bark of trees, but the highest concentration is found in the pericarb of the mangosteen fruit. There are 200 known Xanthones; 44 of them are found in the pericarb of the mangosteen fruit.

Personal communication from Dr. Vaughn T. Johnson, DO told of his open study on patients who had elevated C-Reactive Protein CRP), an inflammatory marker. After putting them on the whole mangosteen fruit puree for one month, 29 out of 30 of the patients' CRPs returned to normal. This suggests that the phytonutrients in the Mangosteen fruit puree containing Xanthones have a potent anti-inflammatory property. Many diseases are caused by inflammation, and if there are substances in nature that will help reduce inflammation, people should consider taking these food-source phytonutrients to help reduce inflammation in order to prevent the development of medical diseases.

Traditional health practitioners also use herbs and spices to treat painful conditions and reduce inflammation. Herbs such as bupleurum, ginseng, licorice, ginger, cat's claw, guggul, curcumin, and bromelain have potent anti-inflammatory action and, hence, have helped patients with painful conditions (Bermejo Benito, 1998).

FUTURE TREND

It is heartening to note that patients and health practitioners are looking at alternative ways to help manage their illnesses, including pain. Combining technology and tradition for pain management is moving in the right direction. Present modern treatment methods have leveled off. Using stronger drugs to control pain is not the way to go. Taking a closer look at what is causing the seemingly increased number of people suffering from aches and pains may help us to find a less invasive and maybe more effective solution to the pain problem.

Inflammation holds the key to the pain problem. If we can reduce the background inflammation by changing our lifestyle, our diet, and nutrition to a more inflammatory neutral living, pain will be similarly reduced to a soft background tingle. More and more people and a growing number of health practitioners are looking at alternative ways to help relieve pain. Using scientific knowledge and modern technology and integrating them with some old fashioned remedies from Mother Nature will see us winning the war against pain.

CONCLUSION

We live in an exciting but challenging time. The evolution of technology is outstripping the slow Darwinian's process of evolution. Things are moving too fast, background noise is too loud, light and color are too dazzlingly bright. The poor body is trying to cope with the massive changes. The body, so to speak, is not ready for the 21st century!

As a consequence of these massive changes, a new group of medical diseases has evolved. Pain, an unpleasant sensation, has grown louder and stronger in modern society. People are relying on stronger drugs, more intensive physical therapy, and the latest technology to combat pain. Technology-driven solutions alone are not necessarily the optimal answer, but a combination of both old and new from the East and West may affect superior pain management strategies.

Holistic medicine is the wave of the present and the future. It is good medicine to "first do no harm" and to encourages patients to be actively involved in the healing process using healing foods and herbs provided by the benevolent hands of Mother Nature. This integrative approach will ultimately lead to better practice of medicine and a superior quality of outcome.

REFERENCES

Author's Clinical Observation & Treatment Protocol (unpublished).

Bagis, S., et al. (2005). Free radicals and antioxidants in primary fibromyalgia: An oxidative stress disorder? *Journal Rheumatology International*, *25*(3), 188–190.

Baxter, G.D. (1994). *Therapeutic lasers—Theory and practice*. Churchill Livingstone

Beri, D., et al. (1988). Effect of dietary restrictions on disease activity in rheumatoid arthritis. *Ann Rheum Dis*, *47*, 69–72.

Bermejo Benito, P., et al. (1998). In vivo and in vitro antiinflammatory activity of saikosaponins. *Life Sciences, 63*(13), 1147–1156.

Carroll, D., Moore, R.A., McQuay, H.J., Fairman, F., Tramer, M., & Leijon, G. (2001). Transcutaneous electrical nerve stimulation (TENS) for chronic pain. *Cochrane Database Syst Rev.*, 2001(3). CD003222.

Childers, N.F., & Margoles, M.S. (1993). An apparent relation of nightshades (solanaceae) to arthritis. *Journal of Neurological and Orthopedic Medical Surgery, 12*, 227–231.

Chinese acupuncture and moxibustion. (1990). Beijing: Foreign Languages Press.

Darlington, L.G., & Ramsey, N.W. (1991). Diets for rheumatoid arthritis. *Lancet*, *338*, 1209.

Darlington, L.G., Ramsey, N.W., & Mansfield, J.R. (1986). Placebocontrolled, blind study of dietary manipulation therapy in rheumatoid arthritis. *Lancet*, *i*, 236–238.

Fargas-Babjak, A., Rooney, P., & Geretz-Simon, E. *Randomised trial of Codetron for pain control in osteoarthritis of the hip & knee*.

Godfrey, C.M. *Antihabituation: A new modality for chronic pain relief.*

Holdcroft, A., & Power, I. (2003). Recent developments: Management of pain. *BMJ*, 326, 635–639.

James, M.J., Gibson, R.A., & Cleland, L.G.. (2000). Dietary polyunsaturated fatty acids and inflammatory mediator production. *Am J Clin Nutr, 71*(1 Suppl), 343S–348S.

KjeldsenKragh, J., et al. (1991). Controlled trial of fasting and oneyear vegetarian diet in rheumatoid arthritis. *Lancet, 338*, 899–902.

Kremer, J.M., et al. (1990). Dietary fish oil and olive oil supplementation in patients with rheumatoid arthritis. Clinical and immunologic effects. *Arthritis Rheum, 33*, 810–820.

Linos, A., et al. (1999). Dietary factors in relation to rheumatoid arthritis: A role for olive oil and cooked vegetables. *Am J Clin Nutr, 70*, 1077–1082.

Lucas, C.P., & Power, L. (1981). Dietary fat aggravates active rheumatoid arthritis. *Clin Res, 29*, 754A.

Milne, S., et al. (2001). Transcutaneous electrical nerve stimulation (TENS) for chronic low back pain. *Cochrane Database Syst Rev.* 2001;(2). CD003008.

Nenonen, M., Helve, T., & Hanninen, O. (1992). Effects of uncooked vegan food—"living food"—on rheumatoid arthritis, a three month controlled and randomised study. *Am J Clin Nutr, 56*, 762.

O'Connor, J., & Bensky, D. (Eds.). (1983). *Acupuncture—A comprehensive text. Shanghai College of Traditional Medicine*. Chicago: Eastland Press.

Oshiro, T., & Calderhead, G. (1990). *Photobioactivation, low level laser therapy: A practical introduction*. A Wiley medical publication.

Packer, L., Hiramatsu, M., & Yoshikawa, M. (1999). *Antioxidant food supplements in human health*. Academic Press.

Panush, R.S. (1988). Possible role of food sensitivity in arthritis. *Ann Allerg, 61*(2), 31–35.

Personal communication with Dr. Vaughn T. Johnson, DO, Utah, USA.

Shir, M.D., Raja, S.N., Weissman, C.S., Campbell, J.N., & D.M.D. Ze'ev Seltzer. (2001). Consumption of soy diet before nerve injury pre-empts the development of neuropathic pain in rats. *Anesthesiology, 95*, 1238–1244.

Skoldstram, L. (1987). Fasting and vegan diet in rheumatoid arthritis. *Scand J Rheumatol, 15*, 219–221.

Taylor, M.R. (1983). Food allergy as an etiological factor in arthropathies: A survey. *J Internat Acad Prev Med, 8*, 28–38.

Warmbrand, M. (1974). *How thousands of my arthritis patients regained their health*. New York: Arco Publishing.

Wong, T.W., & Fung K.P. Acupuncture: From needle to laser. *Family Practice, 8*(2), 168–170.

Zeller, M. (1949). Rheumatoid arthritis—Food allergy as a factor. *Ann Allerg, 7*, 200–205, 239.

KEY TERMS

Antioxidants: Antioxidants are chemical substances that donate an electron to the free radical and convert it to a harmless molecule.

Autoimmune Disease: Normally, the immune system protects us from disease and infection, but if our immune system turns around and mistakenly attacks ourselves, an autoimmune disease ensues. Autoimmune diseases can affect many parts of the body such as connective tissues, nerves, muscles, endocrine system, and digestive system.

Codetron: Codetron has been classified as acupuncture-like TENS. The effect of this machine is equal to or better than electroacupuncture therapy.

Eicosanoid: Eicosanoid is a lipid mediator of inflammation derived from the 20-carbon atom arachidonic acid (20 in Greek is "eicosa") or a similar fatty acid. The eicosanoids include the prostaglandins, prostacyclin, thromboxane, and leukotrienes.

Free Radical: A free radical is an atom or molecule that has an unpaired electron in the outer ring. An unpaired electron will also always mean that there is an odd number, since pairing of electrons goes by 2s. Free radicals are highly reactive molecules and cause damage to nearby cells, causing a cascade of chain reactions of free radical damage to other cells and tissues.

Inflammation: Inflammation is a basic way in which the body reacts to infection, irritation, or other injury, the key feature being redness, warmth, swelling, and pain.

LLLT or Low Level Laser Therapy (also known as Low Power Laser Therapy): LLLT is the application of red and near infrared light over injuries or lesions to improve wound/soft tissue healing and give relief for both acute and chronic pain, nonhealing wounds, post-op pain, and acupoint stimulation. First developed in 1967, it is now commonly referred to as LLLT. The red and near infrared light (600nm-1000nm) commonly used in LLLT can be produced by laser or high-intensity LED. The intensity of LLLT lasers and LEDs is not high like a surgical laser. There is no heating effect.

Nightshade Plants: Solanaceae (nightshade family) has been highly cultivated over the years and includes potatoes, tomatoes, eggplant, sweet peppers, chili peppers (but not black pepper), tobacco, and petunias. Some plants have great medicinal value, and some are quite poisonous. The calyx and corolla are 5-lobed, and usually there are 5 stamens. The fruit is partitioned into 2 or 4 seed-producing divisions, and may be dry or a fleshy berry.

Pygnogenol: Pygnogenol is the patented trade name for a water extract of the bark of the French maritime pine (Pinus pinaster ssp. atlantica). Pygnogenol contains *oligomeric proanthocyanidins* (OPCs) as well as several other Bioflavanoids: *catechin, epicatechin, Phenolic fruit acids,* and *taxifolin*. Procyanidins are oligomeric catechins found in red wine, grapes, cocoa, cranberries, and apples. Pygnogenol helps the integrity of the blood vessels, skin, and mental function, as well as being a potent antioxidant.

Qi (pronounced "chee"): Qi is proposed to regulate a person's spiritual, emotional, mental, and physical balance and to be influenced by the opposing forces of Yin (negative energy) and Yang (positive energy). Disease is proposed to result from the flow of Qi being disrupted and Yin and Yang becoming imbalanced.

TENS (Transcutaneous Electrical Nerve Stimulation). TENS produces neuromodulation through the following pathways: (1) presynaptic inhibition in the dorsal horn of the spinal cord; (2) endogenous pain control (via endorphins, enkaphalins, and dynorphins); (3) direct inhibition of an abnormally excited nerves; and (4) restoration of afferent input. The results of laboratory studies suggest that electrical stimulation delivered by a TENS unit reduces pain through nociceptive inhibition at the presynaptic level in the dorsal horn, thus limiting its central transmission. The electrical stimuli on the skin preferentially activate low-threshold myelinated nerve fibers. The afferent input from these fibers inhibits propagation of nociception carried in the small unmyelinated C fibers by blocking transmission along these fibers to the target or T cells located in the Substantia Gelatinosa (laminae 2 and 3) of the dorsal horn.

Xanthones: Xanthones are close cousins to the polyphenols family and have strong antioxidant effects on the nervous system. They are found in several botanical tonics, including St. John's wort, gentian root nectar, and mangosteen. Current research on Xanthones suggests that they are beneficial in helping with many conditions, including allergies, infections (microbial, fungus, viral), cholesterol levels, inflammation, skin disorders, gastrointestinal disorders, and fatigue. Xanthones have also been found to support and enhance the body's immune system and exhibit strong antioxidant activity, which is beneficial for neutralizing free radicals in the body.

A Comparison of Lossless Image Compression Algorithms for Colour Retina Images

Gerald Schaefer
Aston University, UK

Roman Starosolski
Silesian University of Technology, Poland

INTRODUCTION

Diabetic retinopathy is the leading cause of blindness in the adult population. In order to effectively identify patients suffering from the disease, mass-screening efforts are underway during which digital images of the retina are captured and then assessed by an ophthalmologist. In order to identify features such as exudates and microaneurysms, which are typically very small in extent, retinal images are captured at high resolutions. This in turn means large file sizes and, considering the archival of typically thousands of records, a high demand on computational resources, in particular storage space as well as bandwidth when used in a Picture Archiving and Communications System (PACS). Image compression therefore seems a necessary step. Image compression algorithms can be divided into two groups: lossy techniques where some of the visually less important image data is discarded in order to improve compression ratios, and lossless methods which allow the restoration of the original data. As the features that indicate retinopathy are very small in size and following legislation in several countries, only lossless compression seems suitable for retinal images.

In this article, we present experiments aimed to identify a suitable compression algorithm for colour retina images (Schaefer & Starosolski, 2006). Such an algorithm, in order to prove useful in a real-life PACS, should not only reduce the file size of the images significantly but also has to be fast enough, both for compression and decompression. Furthermore, it should be covered by international standards such as ISO standards and, in particular for medical imaging, the Digital Imaging and Communication in Medicine (DICOM) standard (Mildenberger, Eichelberg, & Martin, 2002; National Electrical Manufacturers Association, 2004). For our study, we therefore selected those compression algorithms that are supported in DICOM, namely TIFF PackBits (Adobe Systems Inc., 1995), Lossless JPEG (Langdon, Gulati, & Seiler, 1992), JPEG-LS (ISO/IEC, 1999), and JPEG2000 (ISO/IEC, 2002). For comparison, we also included CALIC (Wu, 1997), which is often employed for benchmarking compression algorithms. All algorithms were evaluated in terms of compression ratio which describes the reduction of file size and speed. For speed, we consider both the time it takes to encode an image (compression speed) and to decode (decompression speed), as both are relevant within a PACS.

Experiments were performed on a large dataset of more than 800 colour retinal images, which were also divided into subgroups according to retinal region (nasal, posterior, and temporal) and images size. Overall, JPEG-LS was found to be the best performing algorithm as it provides good compression ratios coupled with high speed.

LOSSLESS IMAGE COMPRESSION ALGORITHMS

In this section, we give a brief overview of the lossless compression algorithms that we have evaluated. For further details on the algorithms, we refer the reader to the original references.

- **TIFF PackBits:** A simple RLE compression algorithm (Adobe Systems Inc., 1995). The Tag Image File Format (TIFF) standard specifies this simple runlength (RLE) coding technique; we used the LibTIFF implementation by Leffler (version 3.6.1, http://www.remotesensing.org/libtiff/).
- **Lossless JPEG:** Former JPEG (Joint Photographic Experts Group) committee standard for lossless image compression (Langdon et al., 1992). The standard describes predictive image

compression algorithm with Huffman or arithmetic entropy coder. We used the Cornell University implementation (version 1.0, ftp://ftp.cs.cornell.edu/pub/multimed/ljpg.tar.Z) which applies Huffman coding. The results are reported for the predictor function SV7 which resulted in the best average compression ratio for the dataset.

- **JPEG-LS:** The standard of the JPEG committee for lossless and near-lossless compression of still images (ISO/IEC, 1999). The standard describes low-complexity predictive image compression algorithm with entropy coding using modified Golomb-Rice family. The algorithm is based on the LOCO-I algorithm (Weinberger, Seroussi, & Sapiro, 1996). We used the University of British Columbia implementation (version 2.2, ftp://ftp.netbsd.org/pub/NetBSD/packages/distfiles/jpeg_ls_v2.2.tar.gz).
- **JPEG2000:** A more recent JPEG committee standard describing an algorithm based on wavelet transform image decomposition and arithmetic coding (ISO/IEC, 2002). Apart from lossy and lossless compressing and decompressing of whole images, it delivers many interesting features such as progressive transmission, region of interest coding, and so forth (Christopoulos, Skodras, & Ebrahimi, 2000). We used the JasPer implementation by Adams (version 1.700.0, http://www.ece.uvic.ca/~mdadams/jasper/).
- **CALIC (Context-based Adaptive Lossless Image Compression):** A relatively complex predictive image compression algorithm using arithmetic entropy coding, which because of its usually good compression ratios is commonly used as a reference for other image compression algorithms (Wu, 1997). We used the implementation by Yang. In contrast to the other algorithms, CALIC is designed for grayscale images only. We therefore apply CALIC to each of the individual channels separately.

Lossless JPEG, JPEG-LS, and JPEG2000 are covered by international ISO standards whereas TIFF represents an industry standard. All algorithms except CALIC are incorporated into the medical imaging DICOM (Mildenberger et al., 2002; National Electrical Manufacturers Association, 2004) standard.

RETINAL IMAGE DATASET

A large set of over 800 colour retinal images captured at various ophthalmology centres was used in our experiments. The set is large both with regards to the number of images as well as with regards to the actual sizes of individual images. All images were initially obtained in uncompressed 24-bit RGB format. Images contain between 1.4 and 3.5 millions pixels and hence require between 4 and 10 MB of storage space.

In order to verify whether there are certain image classes that are especially susceptible to compression (or especially hard to compress), we divided the dataset into categories according to the following criteria:

1. **Retinal region:** The whole set is divided into three groups: *nasal*, *posterior*, and *temporal*. Evaluating the compression performance for individual subgroups will highlight whether any algorithms work especially well on any of these categories.
2. **Image size:** Here the images fall mainly into two categories: those images with about 1.4 million pixels (*small*) and those with about 3.4 million pixels (*large*). Compression ratios and compression speed are sometimes dependent on the image size, evaluating the individual size categories will hence confirm whether any such dependency exists for retinal images.

The two criteria are independent of each other (i.e., there exist both *small* and *large* images of all three regions in the dataset) and are hence tested separately. Details on how the images are divided into the individual categories are given in Table 1.

EXPERIMENTAL RESULTS

Experimental results were obtained on a HP Proliant ML350G3 computer equipped with two Intel Xeon 3.06 GHz (512 KB cache memory) processors and Windows 2003 operating system. Single-threaded applications of algorithms used for comparisons were compiled using Intel C++ 8.1 compiler. To minimise effects of the system load and the input-output subsystem performance, the algorithms were run several times; the time of the first run was ignored while the collective time of other runs (executed for at least 1 second, and at least three times) was measured and then averaged. The time measured is the sum of time spent by the processor in application code and in kernel functions called by the application, as reported by the operating system after application execution. The speed of implementations is reported in megabytes (uncompressed) per second (MB/s), where 1 MB = 124 bytes, both for compression and decompression speeds. We note that we actually measure the speed of the specific implementation of the given algorithm on the particular computer system, not the absolute speed of the algorithm itself. The computer system we used, in terms of the processor speed, amount of the installed memory, and so forth, is similar to machines currently employed in PACS. The compression ratios are reported as bitrates, expressed in bits per pixel (bpp) $8e/n$, where e is the size in bytes of the compressed image including the header and n is the number of pixels in the image. We note that smaller bitrates mean better compression and that uncompressed images are stored using 24 bpp.

As mentioned above, all algorithms were evaluated in terms of compression ratio, compression speed, and decompression speed. Results were obtained both for the full dataset and for the individual categories outlined above and are given in Tables 2 to 4. The numbers are calculated as the averages for all images contained in a category; since not all groups contain the same number of images the average results for all images are slightly different from the average over all groups.

Table 2 lists the compression ratio results for all categories. From there we can see immediately that the compression performance of the PackBits methods is very different from those of the other algorithms. Considering that uncompressed images require 24 bits per pixel, the 19.36 bpp achieved by PackBits is fairly feeble yet not surprisingly so, as runlength coding is only suitable for images with large uniform patches to produce reasonably compressible runlengths of pixels. As this is not the case for retinal images (apart from the fairly uniform black background), the achieved compression ratios alone disqualify PackBits as a suitable method for compressing retinal images. Among the remaining algorithms, Lossless JPEG is consistently the worst performing algorithm, which again is expected as it is based on relatively simple predictive coding. The best performing algorithm is CALIC with an average bitrate of 6.60 bpp. That CALIC provides the best

Table 1. Test dataset of retinal images

Category	Number of images	Average size (pixels)
nasal	252	2214864
posterior	301	2203523
temporal	250	2213720
small	468	1383184
large	335	3365689
all	803	2210257

Table 2. Compression ratio results

Category	PackBits	L-JPEG	JPEG-LS	JPEG2000	CALIC
nasal	19.38	8.12	6.83	7.14	6.61
posterior	19.25	8.08	6.80	7.11	6.56
temporal	19.48	8.10	6.84	7.16	6.63
small	18.32	7.92	6.49	6.84	6.25
large	20.78	8.34	7.26	7.53	7.07
all	19.36	8.10	6.82	7.14	6.60

Table 3. Compression speed results

Category	PackBits	L-JPEG	JPEG-LS	JPEG2000	CALIC
nasal	51.2	16.4	14.7	3.3	2.7
posterior	51.4	16.5	14.7	3.2	2.7
temporal	51.0	16.4	14.6	3.3	2.7
small	50.2	16.5	15.3	3.4	2.8
large	52.6	16.4	13.8	3.0	2.6
all	51.2	16.5	14.7	3.3	2.7

Table 4. Decompression speed results

Category	PackBits	L-JPEG	JPEG-LS	JPEG2000	CALIC
nasal	68.7	24.3	14.4	3.1	2.4
posterior	69.1	24.5	14.4	3.0	2.4
temporal	68.4	24.3	14.3	3.1	2.3
small	64.8	24.3	15.0	3.2	2.4
large	74.2	24.5	13.5	2.7	2.2
all	68.7	24.4	14.4	3.1	2.4

compression ratio was to be expected and was indeed the reason for including the algorithm in the evaluation in the first place. However, JPEG-LS performs only slightly worse than CALIC with an average bpp value of 6.82. Somewhat worse than JPEG-LS is the performance of JPEG2000 with a compression ratio of 7.14 bits per pixel.

Looking at the variation of results between the image categories we see that there is very little difference between the compression ratios for the *nasal*, *posterior*, and *temporal* image groups. This suggests that these groups share similar image characteristics. Inspecting the *small* and *large* image categories we can see that all algorithms are more effective for smaller images. The larger images of about 3.4 million pixels are compressed with an on average about 10% higher bitrate compared to the smaller 1.4 megapixel images; the highest differences are observed for PackBits while Lossless JPEG is the algorithm that performs most constant across different image sizes.

We now turn our attention to Table 3 which lists the compression speeds expressed in terms of MB/s of all algorithms. Here, PackBits is clearly the best performing technique affording compression speeds more than three times higher the next best competitor. However, as has been pointed out above, this high speed comes at the expense of an unsatisfyingly high bitrate which rules out the algorithm as being useful in practise. Lossless JPEG is the next fastest algorithm with an average compression speed of 16.5 MB/s which is again due to the relatively simple compression technique. Only slightly slower is JPEG-LS with 14.7 MB/s, which hence provides fast compression coupled with good compression ratios. JPEG2000 and CALIC are much slower than the other algorithms with compressions speeds of only about 3 MB/s which is about five times slower than Lossless JPEG and JPEG-LS. This low speed results from the relatively complex compression paradigms involved which are based on wavelets and complex predictive and arithmetic coding respectively. Looking at the individual image groups, the results across the different region category images are again very uniform while larger images allow higher compression speeds with PackBits but lower speeds with JPEG-LS, JPEG2000, and CALIC.

While the compression speed is obviously of importance at the time of image capture, in a PACS the decompression time might be of higher interest as medical images are typically stored (and hence compressed) only once whereas they are read (and hence decompressed) several times, in particular in the context of medical image retrieval applications (Mueller, Michoux, Bandon, & Geissbuhler, 2004). In Table 4, we therefore list decompression speeds, again expressed in terms of MB/s, for the algorithms on all image categories. Since there is always a certain symmetry in the operations involved in encoding and decoding images, it is not surprising to see that overall the decompression speed results are not very different from the compression speed ones. Lossless JPEG represents a notable exception as for this algorithm the decompression speed is much higher (by about 2/3) than the compression speed. PackBits also decompresses faster than it encodes whereas for the rest of the algorithms encoding and decoding speeds are fairly close to each other. Hence, PackBits is by far the fastest algorithm in terms of decompression speed, followed by Lossless JPEG, JPEG-LS, JPEG2000, and CALIC. Again, there is no noticeable difference between the different region categories while PackBits decompresses larger images faster still, and JPEG-LS, JPEG2000, and CALIC have slightly higher decompression speeds for smaller images.

Integrating all experimental results, that is, compression ratios, compression and decompression speeds, should enable us to judge each algorithm's suitability for the task of being employed as a compression algorithm for colour retinal images in a medical Picture Archiving and Communications System. While TIFF Packbits provides by far the highest encoding and decoding speeds, its performance in terms of compression ratios is also significantly worse than those of all others. The fact that it reduces file sizes only by slightly more than 20% rules it out as an algorithm to be used in practice. CALIC on the other hand provides the best compression rates but is at the same time the slowest algorithm. In fact, both CALIC and JPEG2000 are significantly slower than the rest of the algorithms, both for compression and decompression. This slow decompression speed makes them unsuitable to be employed in PACSs; in addition CALIC is not covered by any standard, in particular not by the DICOM standard. The most suitable algorithm for lossless compression of colour retinal images hence seems to be the JPEG-LS standard. It provides good compression ratios (only slightly worse than CALIC) combined with fast compression and decompression performance. Furthermore, JPEG-LS represents an ISO standard and is included in the DICOM standard for medical imaging. In systems where decompression speed is significantly more important that compression ratio, Lossless JPEG might be an alternative as it decodes faster, yet at the expense of higher bitrates.

FUTURE DIRECTIONS

The presented study clearly indicates the suitability of various compression algorithms for the medical domain. Nevertheless, further insights can be gained by conducting "real-world" evaluation in a large PACS and measuring the retrieval times of various images based on different encodings.

In terms of compression ratios, only little improvements can be expected in the future. The reason for this is that current compression algorithms are very good at identifying and exploiting redundant information in images. Other factors however are likely to be improved further such as compression and decompression speed (though potentially at the expense of a small amount of compressibility). Also, other capabilities such as progressive image decoding or the possibility of midstream content access (Picard, 1994) are likely to prove useful in future applications.

CONCLUSION

We have analysed the performance of several standard lossless image compression algorithms, namely TIFF PackBits, Lossless JPEG, JPEG-LS, and JPEG2000, as well as the nonstandard "benchmark" CALIC algorithm for a large set of colour retinal images. The compression performance was measured in terms of compression ratio, compression speed, and decompression speed. JPEG-LS was found to be the best performing algorithm offering good compression ratios combined with high compression and decompression speed. In addition, apart from being a standard itself, JPEG-LS is also incorporated in the DICOM standard and is hence readily available to be employed in Picture Archiving and Communications Systems.

REFERENCES

Adobe Systems Inc. (1995). TIFF 6.0.1 specification.

Christopoulos, C., Skodras, A., & Ebrahimi, T. (2000). The JPEG2000 still image coding system: An overview. *IEEE Trans. Consumer Electronics, 46*(4), 1103-1127.

ISO/IEC. (1999). *Lossless and near-lossless compression of continuous-tone images - baseline.* ISO/IEC International Standard 14495-1.

ISO/IEC. (2002). *JPEG2000 image coding system: Core coding system*. ISO/IEC International Standard 15444-1.

Langdon, G., Gulati, A., & Seiler, E. (1992). On the JPEG model for lossless image compression. In *2nd Data Compression Conference* (pp. 172-180).

Mildenberger, P., Eichelberg, M., & Martin, E. (2002). Introduction to the DICOM standard. *European Radiology, 12*, 920-927.

Mueller, H., Michoux, N. Bandon, D., & Geissbuhler, A. (2004). A review of content-based image retrieval systems in medical applications - clinical benefits and future directions. *Int. Journal of Medical Informatics, 73*(1), 1-23.

National Electrical Manufacturers Association. (2004). *Digital imaging and communications in medicine (DICOM).* Standards Publication PS 3.1-2004.

Picard, R.W. (1994). *Content access for image/video coding: The fourth criterion*. MIT Media Lab Technical Report 195.

Schaefer, G., & Starosolski, R. (2006). An evaluation of image compression algorithms for colour retinal images. In *Int. Symposium on Biological and Medical Data Analysis*, Springer (LNCS 4345, pp. 358-367).

Weinberger, M.J., Seroussi, G., & Sapiro, G. (1996). The LOCO-I lossless image compression algorithm: Principles and standardization into JPEG-LS. *IEEE Trans. Image Processing, 9*(8), 1309-1324.

Wu, X.L. (1997). Lossless compression of continuous-tone images via context selection, quantization, and modeling. *IEEE Trans. Image Processing, 6*(5), 656-664.

KEY TERMS

Diabetic retinopathy: A common complication of diabetes leading to progressive damage of the eye's retina.

DICOM: A medical imaging standard describing how image information can be exchanged.

Image compression: Reduction of the amount of memory used to store an image.

Lossless compression: Compression from which the original information can be recovered without loss.

Ophthalmology: The science of eye medicine.

Retina: A thin layer of neural cells which lines the inner eyeball.

Screening: Checking for a disease in people without symptoms.

Comparison of Step Length Estimators from Wearable Accelerometer Devices

Diego Alvarez
University of Oviedo, Spain

Rafael C. González
University of Oviedo, Spain

Antonio López
University of Oviedo, Spain

Juan C. Alvarez
University of Oviedo, Spain

INTRODUCTION

Video-based gait analysis, and other measurement devices such as force-plates, electrogoniometers, or electrodes to measure EMG signals are adequate for laboratory studies but are not designed for usability in ambulatory applications.

Conversely, accelerometers provide easily portable systems that supply real-time data, which is why they are widely used in ambulatory diagnostic devices. However, such systems do not provide direct measure of several spatio-temporal parameters of interest such as step length, walking distance, or walking velocity. Instead, they have to be estimated with a mathematical model from indirect sensor measurements. Specifically, in this chapter we are concerned with the accelerometry-based estimation of the step length in straight-line human walking.

BACKGROUND

Gait analysis is frequently made by means of video-based recording of markers placed at end-points in a subset of some body segments. However, such equipment is not easily portable and requires off-line digitizing that can be time-consuming and automated real-time acquisition systems, which can be costly. Because of that, nonspecialist users cannot use these systems in ambulatory applications. Other measurement devices such as force-plates, electrogoniometers, or electrodes to measure EMG signals (Grasso, Zago & Lacquaniti, 2000) are also not designed for usability.

Conversely, accelerometers provide easily portable systems that supply real-time data. In addition, these systems come at a decreased cost when compared to video motion analysis systems, making them easily available to a wide range of clinics. Precision and repeatability characteristics of accelerometry make it adequate for gait analysis (Henriksen, Lund, Moe-Nilssen, Bliddal & Danneskiod-Samsoe, 2004), and it has been widely used recently, mainly in ambulatory diagnostic (Auvinet et al., 2002; Kavanagh, Barrett & Morrison, 2004; Moe-Nilssen, 1998a, 1998b; Najafi, Aminian, Paraschiv-Ionescu, Loew, Bula & Robert, 2003).

For instance, there exist empirical relations between the step length and the maximum and minimum vertical acceleration of the body's center of mass (COM) (Analog-Devices, 2000). More frequently, it is accepted as a linear relation of the step length with the step frequency (Ladetto, Gabaglio & Seeters, 2003). Finally, clinical studies show that an inverted pendulum model can describe the displacement of the body's COM while walking, and the model can be applied to estimate the step length from this displacement (Brandes, Zijlstra, Heikens, van Lummel & Rosenbaum, 2006; Zijlstra & Hof, 1997). All these estimators had in common that they require a previous parametric adjustment or calibration from experimental walking data for every individual. After that, they all provide comparable estimation results, as will be shown in this chapter.

In the following, we compare these common step length estimators reported in the literature. Also, modifications to these estimators are proposed, based on biomechanical considerations. Results show that these modifications lead to improvements of interest over previous methods.

EXPERIMENTAL COMPARISON OF STEP LENGTH ESTIMATORS

Participants

Measurements were taken from a group of four adult men, ages 32 to 38. Height of subjects varied from 168cm to 186 cm. None of them presented vestibular or neurological disorders that could affect the experiments, and all of them gave their signed consent.

The step length was estimated from a three-axial accelerometer placed close to the L3 vertebral position, accepted as a fine approximation of the COM position during normal walking (Moe-Nilssen, 1998b). The device is fixed to the lower lumbar spine with an adjustable corset to avoid movement artifacts.

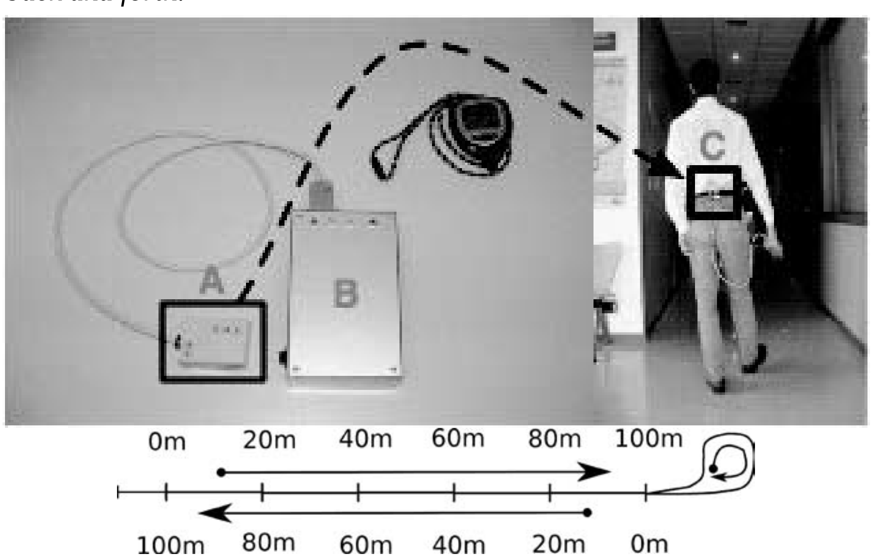
Apparatus

The device prototype (see Figure 1) is built over two biaxial accelerometers AXL202AE from Analog Devices (Analog-Devices, 2000) mounted to form a three-axes frame with a measuring range of ±2*g* being *g* the gravity acceleration. A calibration procedure takes into account offset and scale factors and the orthogonality of the axes (Krohn, Beigl, Decker, Kochendorfer, Robinson & Zimmer, 2005). A PIC16F877 Microchip micro-controller has been used as interface and data logger. Signals are sampled at 50Hz using a 10-bit A/D conversion. Data gathered during the experiments are stored in an internal memory and transferred to an external computer through a serial communication link for further processing. A total of 64K samples can be stored in memory, so we were able to extend the experiments over 20 minutes.

Procedure

During test procedures, subjects were asked to walk along a 100-meter-long corridor following a straight path, back and forth (see Figure 1). The first 10m of walking were discarded for the analysis, as gait is not stable during initial phases of displacement. Each individual completed four independent excursions, and they were asked to maintain a constant velocity for each walk: "preferred" (first excursion), "fast" (second excursion), "low" (third excursion), and "medium" (a velocity between preferred and fast velocities, last

Figure 1. Experimental setup: (Up) Accelerometer device and data acquisition system. The triaxial accelerometer (A) is sampled at 50Hz with the data logger (B) and is fixed to the lower lumbar spine (C) with an adjustable corset (not shown in figure) to avoid movement artifacts. (Down) Procedure: individuals walk a 100m distance at different paces, back and forth.



excursion). This adds up to a total of 800m per individual. Subjects were allowed to turn freely between both 100m segments, and to walk freely or to rest on a chair between different excursions for a lapse of five minutes. Time laps were taken at 20m intervals with a stopwatch. Retest procedures were carried out similarly one month later with the same four subjects and procedure.

Estimators Tested

In its simpler form, human gait can be described by an inverted pendulum model (Zijlstra & Hof, 1997). From this mechanical model, necessary relations between the forward displacement *S* and various measurable step variables can be obtained. Here we will use the relationship between the vertical and the forward displacement, given by equation:

$$S_{M_1} = K2\sqrt{2lh - h^2}$$

where *l* stands for the leg length, and *h* stands for the vertical displacement of the center of mass during one step, as in Figure 2. We well refer to it as the M_1 estimator. The constant *K* has to be calibrated for each individual based on experimental data. The calibration procedure is as follows: for each individual, *K* is the ratio between the real and the estimated walked distance for a given trial.

The vertical displacement for each step is computed with a double integration of the vertical acceleration of the body center of mass. To avoid the integration drift error we need to reset the integral at any point of the step. This reset point was settled at the time of foot-flat, where the vertical velocity of the body is null. Notice that the forward acceleration cannot be used because the COM forward motion has no instants of null velocity, and consequently, the double integration drift error grows unbounded. Mediolateral acceleration, which is responsible of the lateral movement of the hip, presents also zero velocity points, but this signal is weaker than the vertical acceleration and therefore more sensitive to noise.

Estimator M_2 assumes that in the time of foot-flat, the velocity is zero and the vertical coordinate of the COM is the same as at the beginning of the step. Adding a constant offset term to the acceleration forces the final values of the integrals to be the desired ones in the following equations:

$$\int_{t_0}^{t_1} (a_y + c_1)dt = 0$$

$$\int_{t_0}^{t_1} \left(\int_{t_0}^{t} (a_y + c_1)\, dt \right) c_2\, dt = 0$$

being t_0 and t_1 two consecutive instants of foot-flat, and c_1 and c_2 two offsets. This method reduces the drift effects, and a more exact computation of the COM

Figure 2. Inverted pendulum model of human walking. The step length can be estimated as the addition of two factors: the horizontal distance traversed by the COM during the mid-stance phase and the distance traversed during the double-stance phase. Both can be estimated from the vertical maximum displacements of the COM at each respective phase.

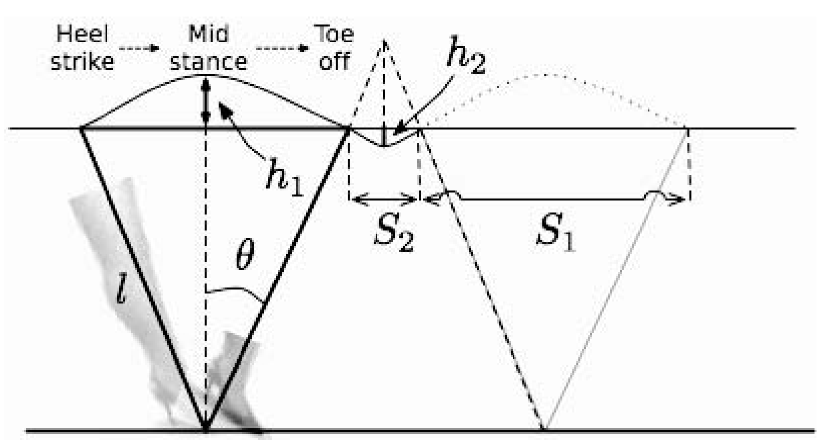
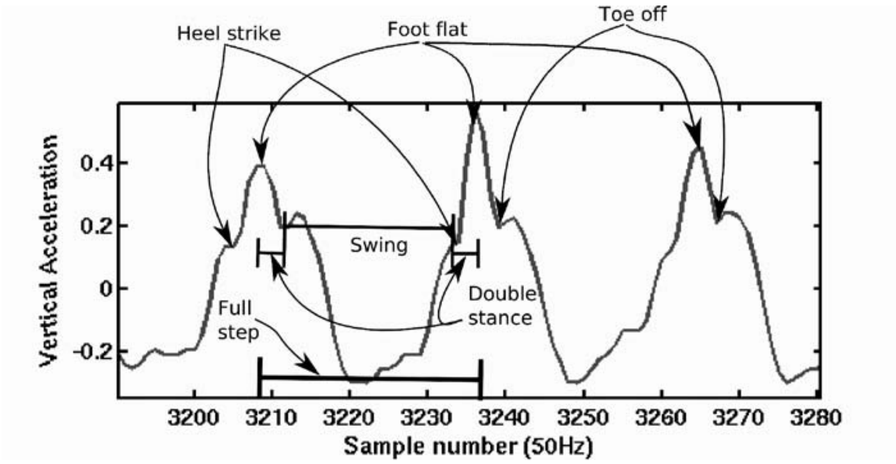


Figure 3. The modified estimator M_3 breaks each step in two phases: double-stance and single-stance. Heel-strike and toe-off events are detected in the measured accelerations. An inverted pendulum model will estimate the longitudinal displacement during the single-stance phase. The displacement during the double-stance phase will be assumed constant.



excursion can be done. This method needs the same calibration as the M_1 estimator.

Estimator M_3 assumes a more complex model (see Figure 3), as the displacement of the COM is related with two pendulums. There is an inverted pendulum with length equal to the length of the leg during the swing phase and a second pendulum (of unknown radius) during the double-stance phase.

$$S_{M_3} = S_1 + S_2 = 2\sqrt{2lh_1 - h_1^2} + S_2$$

The displacement is computed as the motion during the swing phase plus the displacement during the double stance phase. The first displacement can be computed according to the same equation M_1. The vertical excursion of the COM is computed by means of the double integral of the accelerations between the time of toe-off and the time of heel-strike. Detection of toe-off and heel-strike events is done using both the vertical and the antero-posterior accelerations. The same drift correction method used in M_2 is also applied here. The second displacement is set as a constant equal to the foot length, from the first metatarsal head to the calcaneal tuberosity. This method does not need any experimental calibration.

Instead of the double integral, there is an empirical relation of the vertical acceleration with the stride length that depends on the step count *n*:

$$S_{M_4} = Kn\sqrt[4]{a_M - a_m}$$

being a_M and a_m maximum and minimum values of the first harmonic of the vertical accelerations at every step, and *K* a calibration constant that has to be computed from experimental data in the same way as in methods M_1 and M_2. The first harmonic of the acceleration is computed by means of a low pass filter set at 3Hz. This estimator M_4 is claimed to measure distance walked to within 8% across a variety of subjects of different leg lengths (Weinberg, 2002).

The fifth method M_5 is based on the linear relation existing between the cadence and the step length. From experimental data at different speeds, a mean value of cadence and step length is calculated from the first 40m. Those data are fitted to a straight line for each individual. Using this linear model, an estimation of each single step is computed by mean of a linear regression:

$$S_{M_5} = \frac{K_1}{T_{step}} + K_2$$

being T_{step} the period of the step, and K_1 and K_2 two calibration constants.

Results

In Figure 4 there is a comparison among the first four estimators, M_1 to M_4. The vertical axis is the ratio between the estimated and the real walking distance. For each estimator, the most extreme values in the data set (maximum and minimum values), the lower and upper quartiles, and the median are plotted. The estimation has been done for each section of 20m, and the ratio between the estimation and the real ground value (20m) has been computed. The figures show aggregated data for all trials and all individuals. Two calibration procedures have been employed. In the first one (Figure 4-left), *K* was computed using the data from the first 40m walked at the preferred speed. In the second one (Figure 4-right), *K* was computed using the data from the first 40m at each velocity. In both cases, the method M_3 was not experimentally calibrated. Results were computed using the anatomical data of each individual (for methods M_1 to M_3). In Figure 5, all estimators are compared. Every method was calibrated using the data for the first 40m at each speed.

Methods M_1 to M_4 show a tendency to overestimate the step length, as their median values vary between 100% and 105% of the real value. The worst-case error was obtained with method M_1 using single speed calibration, and was almost a 40% higher than the real value. Method M_1 also has the highest variance. When data from multiple speeds are used, all the methods improve, both in median and variance. Method M_5 has the lower median of all (98%).

FUTURE TRENDS

Three methods of step length estimation reported in the literature, M_1, M_2, and M_4, present similar performance. According to the results of our study, there are no significant differences on average among them (see Figure 4). Method M_4 has the advantage that is easier to implement.

Method M_3, which does not need an experimental calibration, achieved similar results to previous estimators that were calibrated with a single velocity procedure (see Figure 4). This has an interesting potential application for wearable devices, as this method avoids unfriendly and time-consuming calibration procedures. Further research is required to define the optimal anatomical landmarks to use as the pendulum length in the swing phase and as the displacement constant during the double stance period.

CONCLUSION

In this chapter, five methods of estimation of step length were compared. Common methods reported in the literature, M_1, M_2, and M_4, present similar performance, with no significant differences on average among the three methods. Method M_1 had more variability for a given prediction. There is no statistical difference between methods M_2 and M_4. These results are consistent with the fact that methods M_2 and M_4 are intended to reduce the effect of the integration drift.

Figure 4. Walking distance estimation, using methods M_1 to M_4 compared to the real walking distance. For each estimator, the distribution (median, quartiles, and extreme values) is represented. (Left) Calibration made using data from the experiment at preferred speed. (Right) Calibration made using data from experiments at different speeds.

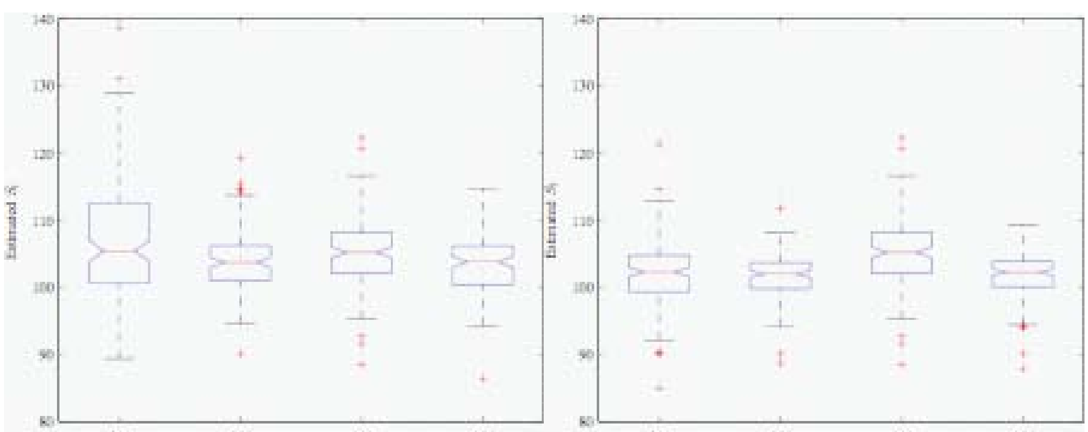
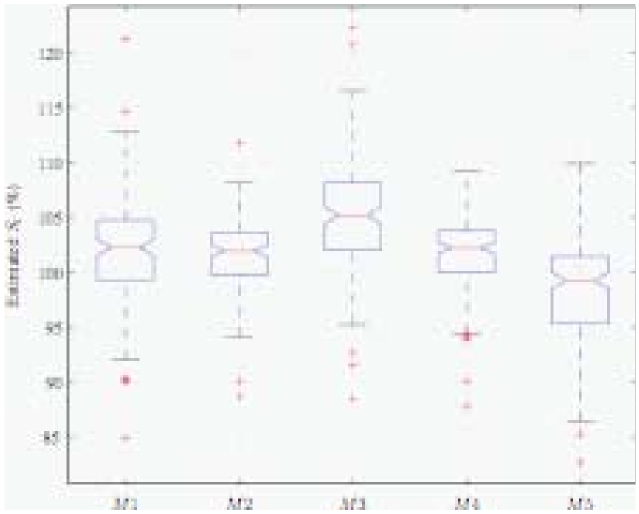


Figure 5. Estimation based on cadence requires data gathered at different velocities for calibration. The figure represents estimated to real walking distance, including the estimator M_5. The new calibration data set is used with all estimators.



Multiple velocity calibration improves the performance of methods M_1, M_2, and M_4, mainly through a significant reduction in the standard deviation (see Figure 4). Median values are closer to 100% (unbiased estimation), compared to single velocity calibration, although the difference is not statistically significant. This result is consistent with the fact that walking speed is a relevant factor of the gait.

Method M_1 produces a closer estimation to the desired value, although it also presents a high variance. A possible reason for this is the low temporal resolution in the measurement of the step duration. Further experiments using a higher acquisition rate are needed to verify this hypothesis. Method M_3, without experimental calibration, achieved similar results to methods M_2 and M_4 calibrated with a single velocity.

ACKNOWLEDGMENT

This work was supported in part by the EU, European Regional Development Fund (ERDF), and by the Spanish Government, grant MEC-05-DPI2005-08573.

REFERENCES

Analog-Devices. (2000). Data sheet, ADXL2002E two axis accelerometer. Analog Devices.

Auvinet, B., et al. (2002). Reference data for normal subjects obtained with an accelerometric device. *Gait & Posture, 16*(2), 124–134.

Brandes, M., Zijlstra, W., Heikens, S., van Lummel, R., & Rosenbaum, D. (2006). Accelerometry based assessment of gait parameters in children. *Gait & Posture* [in press].

Grasso, R., Zago, M., & Lacquaniti, F. (2000). Interactions between posture and locomotion: Motor patterns in humans walking with bent posture versus erect posture. *J Neurophysiol, 83*(1), 288–300.

Henriksen, M., Lund, H., Moe-Nilssen, R., Bliddal, H., & Danneskiod-Samsoe, B. (2004). Test-retest reliability of trunk accelerometric gait analysis. *Gait & Posture, 19*(3), 288–297.

Kavanagh, J., Barrett, R., & Morrison, S. (2004). Upper body accelerations during walking in healthy young and elderly men. *Gait & Posture, 20*(3), 291–298.

Krohn, A., Beigl, M., Decker, C., Kochendorfer, U., Robinson, P., & Zimmer, T. (2005). Inexpensive and automatic calibration for acceleration sensors. In *Proceedings of the Ubiquitous Computing Systems Conference*.

Ladetto, Q., Gabaglio, V., & Seeters, J.V. (2003). U.S. Patent No. 09/839,204.

Moe-Nilssen, R. (1998a). A new method for evaluating motor control in gait under real-life environmental conditions. Part 1: The instrument. *Clinical Biomechanics, 13*(4-5), 320–327.

Moe-Nilssen, R. (1998b). A new method for evaluating motor control in gait under real-life environmental conditions. Part 2: Gait analysis. *Clinical Biomechanics, 13*(4-5), 328–335.

Najafi, B., Aminian, K., Paraschiv-Ionescu, A., Loew, F., Bula, C., & Robert, P. (2003). Ambulatory system for human motion analysis using a kinematic sensor: Monitoring of daily physical activity in the elderly. *IEEE Transactions on Biomedical Engineering, 50*(6), 711–723.

Weinberg, H. (2002). Using the ADXL202 in pedometer and personal navigation applications analog devices, application note.

Zijlstra, W., & Hof, A.L. (1997). Displacement of the pelvis during human walking: Experimental data and model predictions. *Gait & Posture, 6*(3), 249.

KEY TERMS

Ambulatory Inertial Systems: Wearable devices based on inertial sensors.

Inertial Gait Analysis: The process of quantification and interpretation of animal locomotion, including human, from inertial sensors.

Inertial Step Length Estimators: Algorithms that compute the displacement of an artifact based on the signals provided in real time by onboard inertial sensors.

A Complex Non-Contact Bio-Instrumental System

Dan-Marius Dobrea
Technical University "Gh. Asachi," Romania

Adriana Sîrbu
Technical University "Gh. Asachi," Romania

INTRODUCTION

One of the major challenges that the human computer interface (HCI) faces nowadays is that of identifying a subject's state, in a real world environment, characterized mainly by its being: open-recorded, event-elicited, and internal *emotional state*-driven (Picard, Vyzas, & Healey , 2001). The main requirement for such system regards the noninvasive character of their working principle.

Subsequently, in order to improve communication in HCI systems or to asses the human state, the analysis of the *body language* could be a solution. Thus, a "sensitive computer" could use the body movements and the positions of the body in order to assess the state of a person (e.g., confusion, illness, nervousness, lack of attention, motor fatigue, agitation, etc.).

In the rehabilitation process, the measurements of the motion impairments are very important because they can quantify the patient's recovering between two consecutive medical sessions. Nowadays, this type of motion analysis is achieved by physicians through visual observation of the patient during some standard tests. As a result, the physician subjectivism is introduced and, much more, when different physicians evaluate the same patient, the reproducibility of the measurements becomes a difficult task.

To respond to the previously presented requirements in different application fields, a noncontact laser system was introduced by the authors (Cracan, Teodoru, & Dobrea, 2005; Dobrea, 2002; Dobrea, Cracan & Teodoru, 2005; Dobrea & Serban, 2005). Here, we present the implementation of an independent system constructed as a self-contained unit that can be further integrated in much more complex and intelligent structures, together with new possible applications.

This article proposes a new real-time, noncontact system able to:

- Acquire and interpret the subject's *body language*,
- Recognize static *hand signs*, and
- Provide physicians with a quantitative tool to monitor the evolution of the *Parkinson Disease.*

BACKGROUND

The proposed bio-instrumental system (BIS) was designed to be used in the medical field, in applications such as: rehabilitation, functional movement analysis, evaluation of the cognitive deficits, or motion and support offered to the vocally impaired subjects.

Nowadays, in order to evaluate and assess the severity of *Parkinson's Disease*, physicians use different rating scales. The method used to assess *Parkinson's Disease* is based on a questionnaire : Unified Parkinson's Disease Rating Scale, (MDSTF, 2003). The most important disadvantage of the rating scales is the lack of results reproducibility. Different physicians obtain different results on the same patient due to different medical experience and the possibility to observe, at one moment, only one cross-section of the patient. The BIS presented in this article will be used in the quantitative analysis of the head tremor movements. Even if, for this application, other methods exist to acquire the movement, based on accelerometer sensors (Keijsers, Horstink, & Gielen, 2003), optical data flow, and gyroscope (Mayagoitia, Nene, & Veltink, 2002), no method has imposed yet as a standard.

Recognition of the *hand signs* is a challenging task for the nowadays systems and it is very important for the vocally impaired people. Even if the research in this field fade in time, the first large recognized device for identifying the *hand signs* was developed by Dr. G. Grimes (1983) at AT&T Bell Labs. This device was created for "alpha-numeric" characters communication by examining hand positions like an alternative tool to keyboards; it was also proved to be effective as a tool for allowing nonvocal users to "finger-spell" words and phrases. In order to understand the *hand signs* language the hand gesture must be acquired. Mainly, the *hand signs* are acquired using video cameras (Cui & Wenig, 1999; Ho, Yamada, & Umetani, 2005) or some devices that directly determine the position of the hand parts (such as gloves) (Hernandez-Rebollar, Kyriakopoulos, & Lindeman, 2004).

There are strong relations between psychological states and the body movements, confirmed by the theories of Kestenberg-Amighi, Loman, Lewis, and Sossin (1999) and Hunt (1968) or by the analyses realized in the field of the *body language* investigation (Pease, 1992). Moreover, these relationships make the subject of the somatic theory. The health care efficiency in the activity related to the human-computer interaction is directly dependent on both the subject's state and the capability of the health care systems to recognize the specific needs of the user in order to change their response accordingly. Unfortunately, acquiring and interpreting this kind of information is very difficult and, as a consequence, all the actual systems have only a limited ability of communication. Current strategies for user's *emotional state* acquisition are either obtrusive (Picard et al., 2001) or the data captured by the systems consist in low level useful information.

A NEW TYPE OF NONCONTACT BIO-INSTRUMENTAL SYSTEM

The new proposed BIS was designed to determine, in a fast way and without any physical contact with the subject, the movements, the position, and the distance to an observation point. Using this information, the physiological and *emotional states* of the subject are estimated.

System's Architecture. Working Principle

The BIS is composed of a laser scanner, an interface unit, a video camera, and a software program, running on a DSP platform that controls the scanner, acquires the images, and extracts the distance/position information, as in section a of Figure 1. The BIS schematics and the system data flow are presented in section b of Figure 1.

The working principle of the whole system is based on a laser scanner that generates a laser plane at a constant angle from the horizontal plane. When the laser plane hits a target in the imaged area, a line of laser light appears on the body of the subject. See Figure 2 – Img_{t+1} image. The video camera acquires two images: first, with the laser diode off, Img_t, and second, with the laser diode on, with a line of laser light that appears on the target, Img_{t+1}. Subtracting the two images we get only the laser line projected on the people's torso, *OutImg* – Figure 2. In the ideal situation, all pixels for which $Img_{t+1}(x, y) \neq Img_t(x, y)$ describe the laser line which appears on the user's body torso. In real cases, the images are corrupted by noise. This problem was solved using an experimentally obtained noise model, σ. The criterion to extract the line of the laser light becomes now: $Img_{t+1}(x, y) – Img_t(x, y) > \sigma$. Other problems, such as shadows, slight body subject movements, light sources, video camera saturation, and background changes do not affect the reliability of the laser line feature extraction. This is happening because the time interval between the two images acquisition is less than 40 ms and the noise model presented above have been proven to be adequate. Based on this operating principle, the extraction of the laser line becomes a very fast task—a major advantage of this system.

If the object is far away, the extracted laser line will be farther from the bottom of the image, h_1. In the opposite situation, it will be closer to the bottom part of the resulting image, h_2. At this point, one knows the angle between the laser scanner and the horizontal plane, the position in space of the video camera and the extracted shape of the laser line on the subject body. The depth information of each point on the extracted laser line is calculated using some basic geometric formulae. Further on, having all these values, we exactly determine the real 3D subject body position with respect to the camera.

The hardware system has two components: the electro-mechanical scanner and the DSP system. The scanner has a low-power laser diode and a mechanical system with mirrors, section a of Figure 1 (Dobrea, 2002). The plate with mirrors is attached to an engine shaft. The DSP system interfaces with the engine control system only through a single digital line that can start/stop the engine.

Since this application deals with images and all these type of applications are considered data and computing-intensive, the TMS320C6416 DSP was chosen due to its: high computing power, large on-chip memory, and efficient data transfer mechanism.

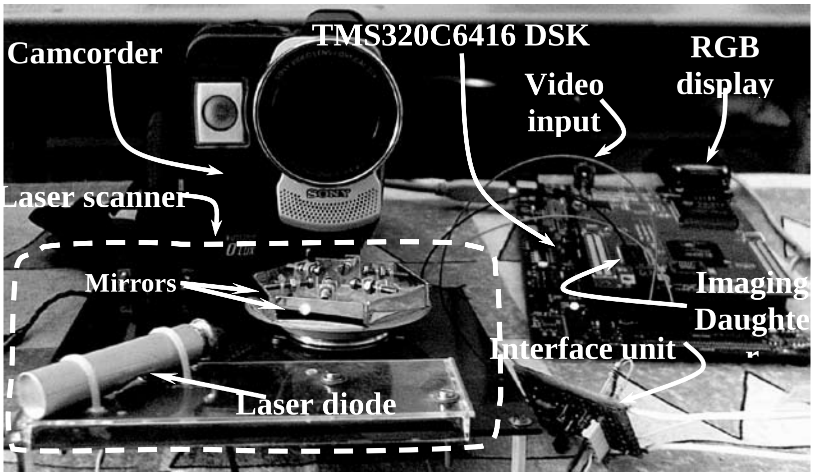
In order to have a *real time supervision* of the system evolution, an output image, containing both the acquired image and the resulting one (*OutImg*), is formed and displayed on a RGB monitor (the image data are moved in background using for this the EDMA controller, section b of Figure 1).

Movement-Based Subject State Identification System

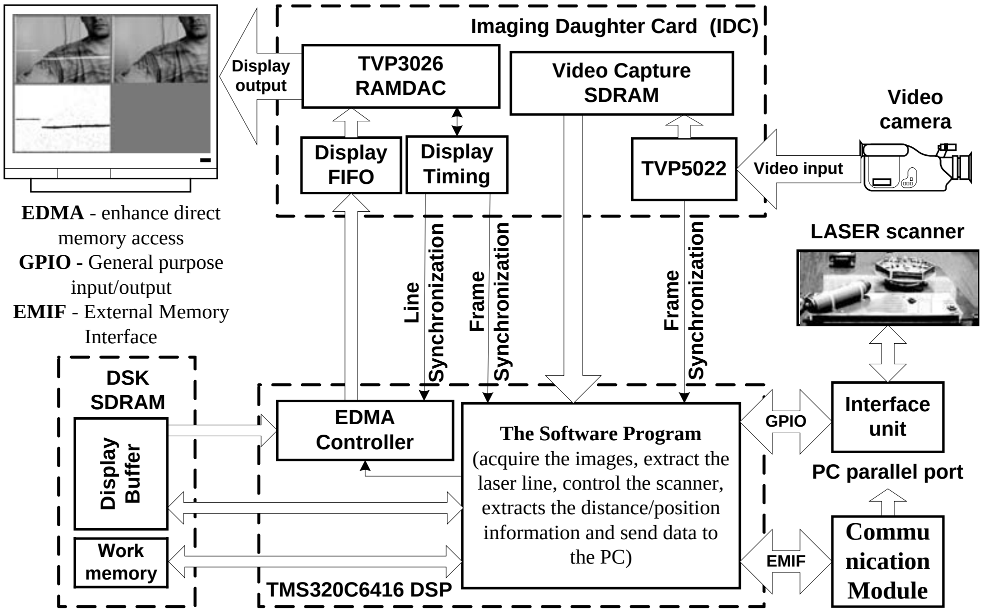
In order to test the BIS, we have developed an experiment intended to determine if there is a correlation between the *emotional state* of a person and the body torso movement of that person.

We admitted six subjects for this study. All of them were young healthy people (26.6±3 years, mean

Figure 1. The bio-instrumental system



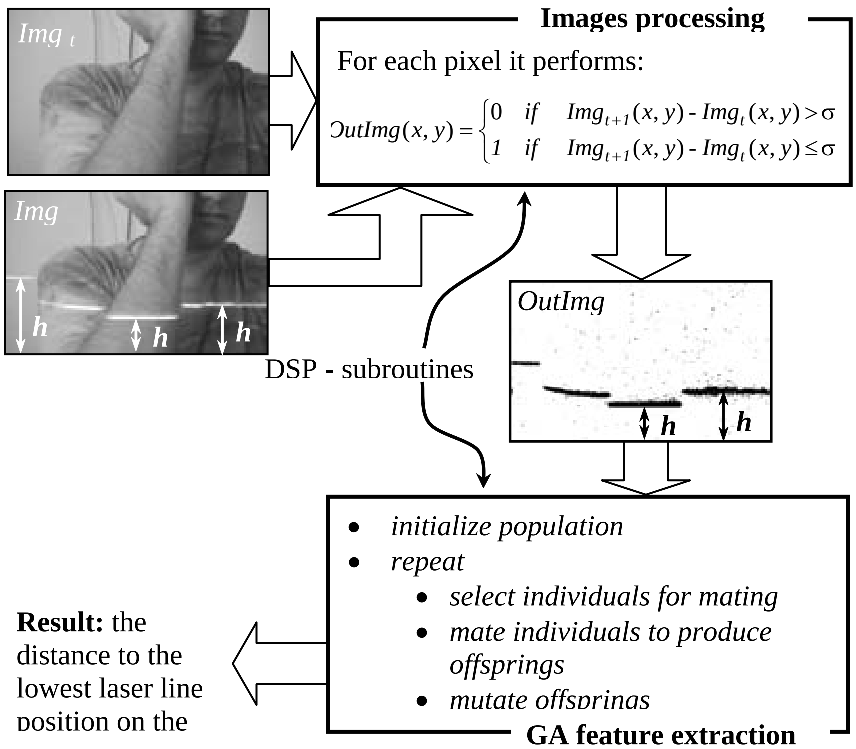
a) view of the implementation



b) BIS schematics and the system data flow

continued on following page

Figure 2. The data flow for the distance determination



± standard deviation) (Dobrea & Serban, 2005). But first, the *emotional state* must exist and must be manifested by the subjects. The emotion was induced by two films presented to the subjects: an action movie and a horror movie. At the end, an analysis was done on the recorded body torso movements to characterize common behaviors of the subjects during the movies associated with special time moments of the films. In this way, the system was validated and analyzed.

The movement of the subject was characterized by the position of the subject torso, determined by means of the distance between the closest point of the chest situated on the laser line and the video camera. This distance is proportional to the distance from the lowest point of the extracted laser line (projected on the subject torso) to the bottom border of the image, h_3 on Figure 2.

The Distance Determination

A special algorithm was designed in order to determine the distance between the subject and the laser diode. The algorithm use a standard genetic algorithm (GA), described by Goldberg (1989), Figure 2. For each generation, an entirely new population is created by selecting individuals for mating from the previous population, according to a specified selection method. In order to implement the GA we have adopted a philosophy inspired by GALib, a C++ library for GA objects (Wall, n.d.).

For our application, each chromosome has two genes, encoding the position of an image pixel (Dobrea, Sîrbu, & Serban, 2004). We have implemented the binary string format for chromosomes, concatenating the binary representations of the coordinates, on the *x* and *y* axis. The fitness function was designed to maximize the number of image pixels in the vicinity of the selected image pixel and to minimize the distance between the pixel and the bottom of the screen. In this mode, the chromosome with the best fitness value will characterize a point belonging to the laser line segment which is closest to the bottom border of the image. To reduce the influence of noise, the pixels having in their vicinity less than a specified amount of adjacent pixels are ignored. The population was initialized randomly, uniformly distributed in the four quadrants of the image, to ensure the rapid convergence of the GA.

In section a of Figure 3, we present two evolutions of the GA for an extracted laser line, displaying the fitness of the best individual and stressing two behaviors: medium and slow algorithm convergence. For the tests we have run, the mean number of generation for convergence is 200, for a population of 100 individuals, with no elitism, 0.9 probability of crossover, and 0.001 probability of mutation.

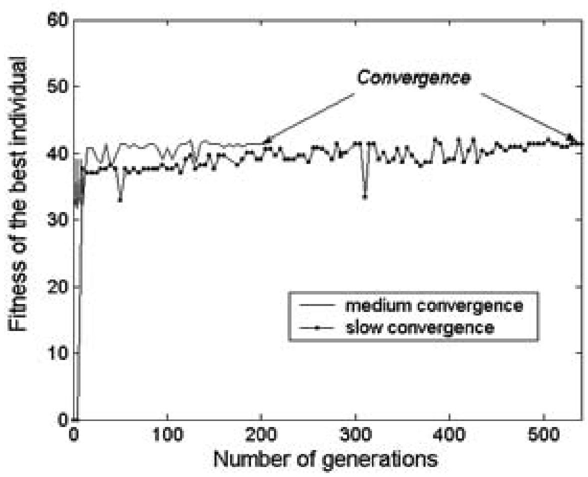
The Emotional State Detection

A problem this system must to deal with is given by the limbs movements that generate *artifacts*—for example, in Figure 2 the arm positioned in front of the body determines the GA to obtain the h_2 distance instead the correct distance h_3. These *artifacts* were removed using a special algorithm which takes into account the arm thickness.

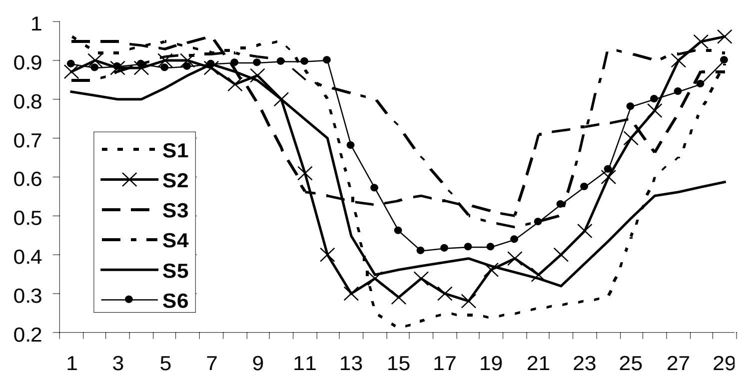
The Pearson's Product-Moment *Correlation coefficient* was computed in order to characterize common behaviors of the subject's movements recorded during the movies and associated with special time events of the films. A time evolution of the distance between the view point (video camera) and the subject's chest position is presented in section b of Figure 3.

The six subject's traces, representing distance evolutions, are presented and marked in this figure with S1- S6. They correspond to a movie fragment able to impress the subjects.

Figure 3. Outcome of the algorithm which determines the distance between the subject and the laser diode



a) Two representative evolutions of the GA



b) The evolution of the torso position for the same segment of the movie for all the subjects

continued on following page

Figure 3. continued

	S1	S2	S3	S4	S5	S6
S1	1	0.852	0.724	0.751	0.960	0.952
S2	0.852	1	0.871	0.781	0.807	0.865
S3	0.724	0.871	1	0.605	0.720	0.710
S4	0.751	0.781	0.605	1	0.553	0.864
S5	0.960	0.807	0.720	0.553	1	0.894
S6	0.952	0.865	0.710	0.864	0.894	

c) The correlation coefficient

The Pearson's *correlation coefficients* computed for all pairs of two time segments, shown in section b of Figure 3, are presented in section c of Figure 3.

The obtained results support and demonstrate the system's ability to evidence a common subject *emotional state* reflected by the body movements. The subjects' different behavior as a response to the same emotional state (through the movements of the body, hands, etc.) and the time delay required to manifest the emotional state determine the spread of the computed correlation values. For other time fragment of the movies, similar results were obtained.

A Hand Sign Recognition System

The system able to recognize the static *hand sign* we propose is a combination of two methods described in the literature: video and special device-based. The *hand signs* can be formed using one or both hands. Five of the ten used *hand signs* are presented in section a and b of Figure 4.

The algorithm used for the extractions of the projected laser line image is similar to the one presented above. The laser trace signal ($H_s[n]$ resulted from the laser extraction algorithm) is modeled using the coefficients of an autoregressive (AR) filter, (Cracan et al., 2005), (Dobrea et al., 2005). The AR filter's coefficients are used to reduce the redundant input information passed to the classifier algorithm implemented on DSP.

A multilayer perceptron (MLP) neural network was used in the *pattern recognition* process (Cracan et al., 2005); Dobrea et al., 2005). The correct recognition rates for all the *hand signs* were in the range of 0.823÷1. The necessary time between the first image acquisition and the end of the entire classification process was less than 1.5 seconds, adequate for *real time supervision.*

FUTURE TRENDS

From the Subject's Body Language to Emotional State Identification

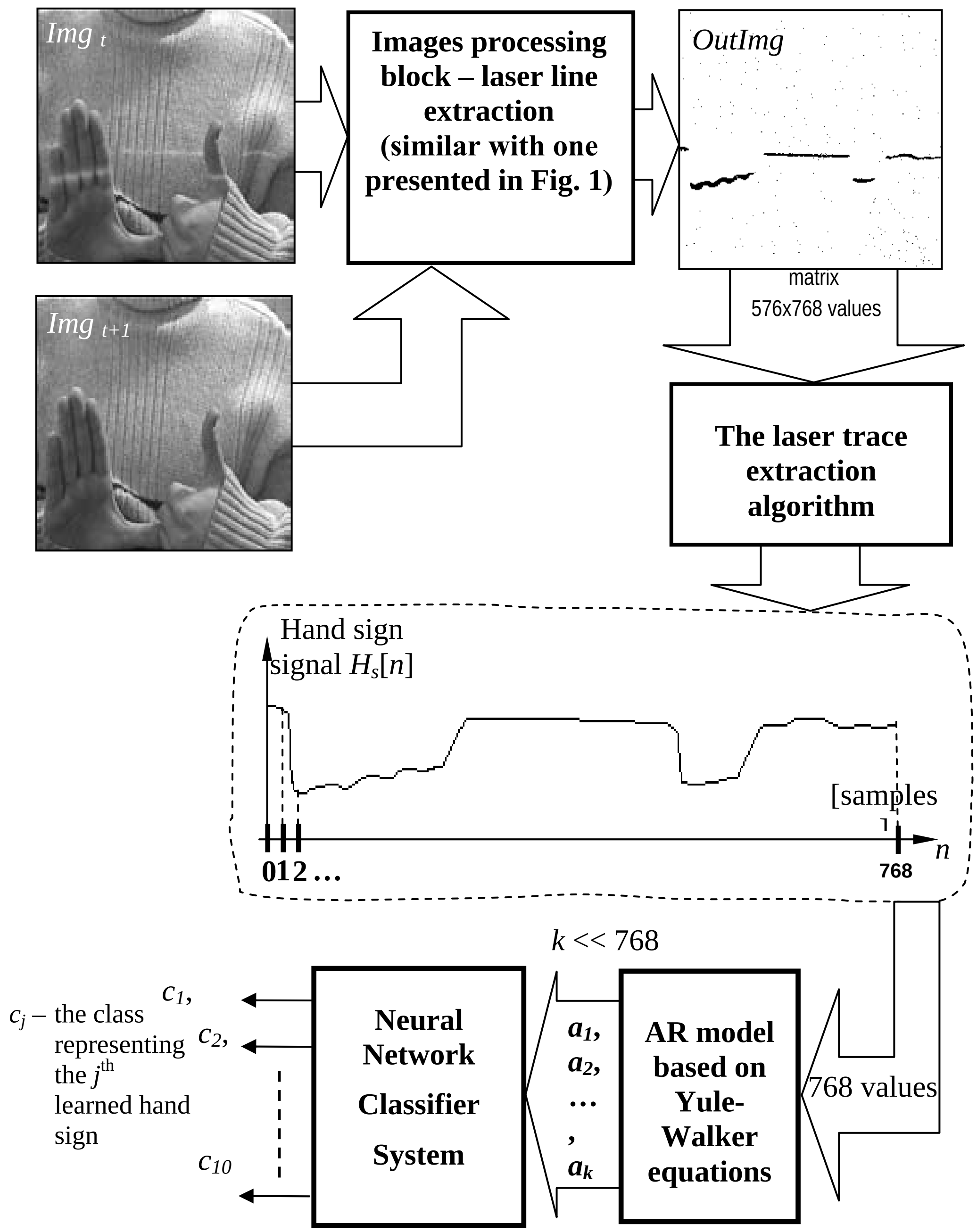
The identification of some particular postures like the arm position in front of the torso, as in Figure 2, or the torso position, as in section a of Figure 5, can be made by *pattern recognition*, in the manner presented above. Each of these postures or body positions can be related to different internal subject states, (Pease, 1992), that can guide a system in order to improve the human computer interaction. For example, in Figure 2, the subject posture can express boring if the subject keeps this posture for a long time.

Evaluation/Analysis of Parkinson Patients

Up until this moment there is no kind of standard method (either qualitative or quantitative) to evaluate the Parkinson symptoms. Moreover, in MDSTF (2003), one mentions a number of errors in the Unified Parkinson's Disease Rating Scale such as: some ambiguity in the text, inadequate instructions to rate some questioner rubrics, one deficiency in a unit of measurement, and the lack of questions other than motor aspects of Parkinson's disease.

Using two different laser scanner systems, the trajectory of the subject head (as in section b of Figure 5) can be recorded and easily quantified in order to assess the patient rehabilitation. In this mode, the proposed system is able to quantitatively evaluate the severity and the progress of the Parkinson's disease and to offer a reproducibility of the obtained results. Thus, all the above presented drawbacks are eliminated.

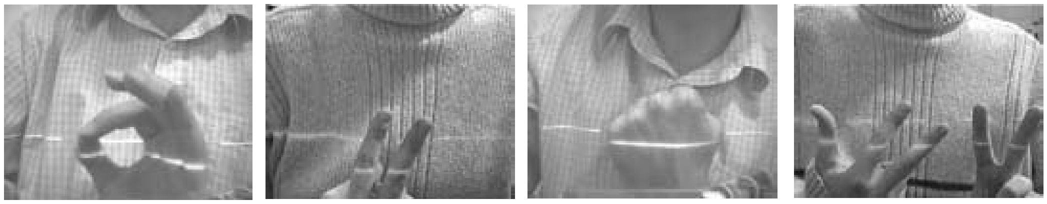
Figure 4. Demonstration of system's ability to evidence subject's emotional state reflected by the body movements



a) The system's flow chart and some of the partial results

continued on following page

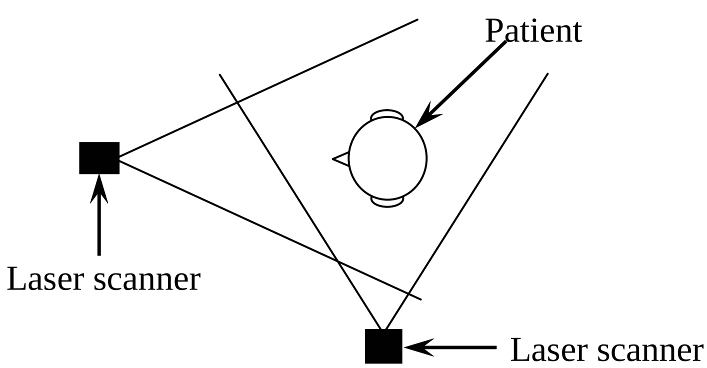
Figure 4. continued



b) Several hand signs recognized by the intelligent system

Figure 5. Identification of some particular postures or movements

a) Image representing a particular body posture



b) The configuration of the laser systems in order to acquire the head position

CONCLUSION

In this article, a DSP implementation of a new noninvasive BIS was presented. This project has a significant impact on the people's life reflected in:

- The natural form of subject's interaction and supervision by the health care systems, in order to determine the emotional and physiological state changing;
- The reproducibility of the evaluation and assessment of the severity in *Parkinson's Disease*—a way of helping the physicians to improve the quality of the medical act;
- The support offered to the vocally impaired subjects.

This system offers the possibility to use a new kind of information regarding the subject's emotional and physiological state, unexploited yet on the HCI systems, namely, the state of the subject expressed through *body language*.

The system is inexpensive, easy to manufacture and, hence, attractive for practical applications. In the end, last but not least, the entire system is very fast, being adequate even for *real time supervision*.

REFERENCES

Cracan, A., Teodoru, C., & Dobrea, D.M. (2005, November 21-24). Techniques to implement an embedded laser sensor for pattern recognition. In *Proceedings of the International Conference on "Computer as a tool,"* Belgrade, Serbia & Montenegro (Vol., 2, pp. 417-1420).

Cui, Y., & Wenig, J. (1999). A learning-based prediction-and-verification segmentation scheme for hand sign image sequence. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 21*, 798-804.

Dobrea, D.M. (2002, December 4-8). A new type of sensor to monitor the body torso movements without physical contact. In *Proceedings of Second European Medical and Biological Engineering Conference*, Austria (Vol., 3, pp. 810-811).

Dobrea, D.M., Cracan, A., & Teodoru, C. (2005, November 20-25). A pattern recognition system for a new laser sensor. In *Proceedings of the 3rd European Medical and Biological Engineering Conference*, Prague, Czech Republic (Vol. 11, pp. 3011-3014).

Dobrea, D.M., & Serban, M.C. (2005, September 14-17). From the movement to emotional state identification. In *Proceedings of the 14th International Conference of Medical Physics*, Nuremberg, Germany (pp. 776-777).

Dobrea, D.M., Sîrbu, A., & Serban, M.C. (2004, June 25-27). DSP implementation of a new type of bioinstrumental noncontact sensor. In *CD Proceedings of 4th European Symposium in Biomedical Engineering*, Patras, Greece.

Goldberg, D. (1989). *Genetic algorithms in search optimization and machine learning*. Addison-Wesley.

Grimes, G.J. (1983, November 8). *Digital data entry glove interface device*. *U.S. Patent No. 4,414,537.* Washington, DC: U.S. Patent and Trademark Office. Bell Telephone Lab. Inc

Hernandez-Rebollar, J.L., Kyriakopoulos, N., & Lindeman, R.W. (2004). A new instrumented approach for translating American sign language into sound and text. In *Proceedings of Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, Seoul, Korea (pp. 547-552).

Ho, M.A.T., Yamada, Y., & Umetani, Y. (2005). An adaptive visual attentive tracker for human communicational behaviors using HMM-based TD learning with new state distinction capability. In *IEEE Transactions on Robotics* (Vol. 3, pp. 497-504).

Hunt, V. (1968). *The biological organization of man to move*. Impulse.

Keijsers, N.L.W., Horstink, M.W.I.M., & Gielen, S.C.A.M (2003). Online monitoring of dyskinesia in patients with Parkinson's Disease. *IEEE Engineering in Medicine and Biology*, *22*(3), 96-103.

Kestenberg-Amighi, J., Loman, S., Lewis, P., & Sossin, S. (1999). *The meaning of movement*. Gordon & Breach Publishers.

Mayagoitia, R.E., Nene, A.V., & Veltink, P.H. (2002). Accelerometer and rate gyroscope measurement of kinematics: An inexpensive alternative to optical motion analysis systems. *Journal of Biomechanics*, *35*(4), 537-542.

Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease. (2003). The unified Parkinson's Disease rating scale (UPDRS): Status and recommendations. *Movement Disorders, 18*(7), 738-750.

Pease, A. (1992). *Body language – how to read other's thoughts by their gesture*. London: Sheldon Press.

Picard, R.W., Vyzas, E., & Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(10), 1175-1191.

Wall, M. (n.d.). GALib. Retrieved February 15, 2008, from http://lancet.mit.edu/ga

KEY TERMS

Biosensor: A device incorporating a biological sensing element (the recognition element is biological in nature).

Digital Signal Processor (DSP): A specialized, programmable computer processing unit that is able to perform high-speed mathematical processing. DSP refers to manipulating analogue information that has been converted into a digital (numerical) form.

Fitness: A measure of the suitability of a potential solution for the given application. Each individual is an encoded representation of all the parameters that characterize the solution. It has an associated value (fitness) which is a measure of its performances.

Genetic Algorithm (GA): An optimization and search technique based on the principles of Darwin's theory of natural selection and Mendel's work in genetics on inheritance: the stronger individuals are likely to survive in a competing environment. It allows a population composed of many individuals (possible solutions) to evolve under specified selection rules to a state that maximizes their suitability for the specific application.

Human Computer Interface (HCI): The process by which users interact with computers; based on HCI, one designs and implements human-centric interactive computer systems.

Neural Network: A system of programs and data structures that approximates the operation of the human brain.

Sensor: A device that measures or detects a real-world condition, such as an acoustic (microphone, hydrophone), electromagnetic (radar), or optical (camera) signal.

C

Computational Intelligence and Sensor Networks for Biomedical Systems

Daniel T. H. Lai
Melbourne University, Australia

Jussi Pakkanen
Helsinki University of Technology, Finland

Rezaul Begg
Victoria University, Australia

Marimuthu Palaniswami
Melbourne University, Australia

INTRODUCTION

The growing demand for a better quality of life is now the major motivation behind the recent focus on health care research. Current worldwide shortages in medical personnel have placed immense strain on present health care systems in the presence of a growing global population. Alarming statistics from the recent WHO 2006 report[1] depict critical shortages in health care providers with 57 countries possessing less than 80% coverage of health care providers for their citizens. It is estimated that training the health care workforce to cope with these shortages by 2015 would incur annual costs of \$1.6 to \$2.0 billion per country depending on its size. This predicament makes it paramount to look at alternative avenues for more effective health care technologies to ensure the sustainability of worldwide health care systems.

Computational intelligence (CI) is the implementation of artificial intelligence in computer science allowing for the design of powerful decision systems capable of processing and interpreting large volumes of data. This discipline has recently been applied to biomedical engineering problems, specifically those in which human intervention can be replaced by some form of automated response. As will be seen later, supervised learning formulations such as artificial neural networks (Haykin, 1994), support vector machines (Schlkopf, Burges, & Smola, 1999; Vapnik, 2000), and fuzzy classifiers (Abe, 1997) have been recently applied to diagnostic and prognostic biomedical applications. These techniques learn nonlinear relationships between patient data and disorders which could not otherwise be captured with standard statistical analysis techniques. Moreover, CI techniques are cost effective to implement, being fundamentally software-based in nature and thus requiring only a processor with mathematical functions and adequate memory, for example, personal computers.

Sensor networks (SN) is an emergent technology which combines small sensors outfitted with wireless transmitters to form a network with more powerful sensing capabilities (Akyildiz, Su, Sankarasubramaniam, & Cayirci, 2002; Chong & Kumar, 2003). The primary application for SN technology is monitoring environmental changes making it ideal for deployment in patient monitoring systems. In contrast to other monitoring technologies such as video, SN offers a potentially cheaper solution consisting of cost effective interconnected sensors which cooperatively sense the surroundings. Individual sensor information is then fused to derive an instantaneous description of the environment.

In this article, we review briefly the recent applications of CI and SN technologies in health care, mentioning some of the challenges in deploying these technologies. This is followed by an example of a biomedical system incorporating both technologies in a single paradigm. The state of current systems and their advantages over existing methods are highlighted with examples focusing primarily on intelligent automated diagnostic systems to augment clinician diagnoses and health care monitoring systems for continuous patient observation.

COMPUTATIONAL INTELLIGENCE FOR DIAGNOSIS AND DETECTION IN HEALTH CARE

In recent times, intelligent biomedical applications incorporating CI techniques have been investigated for detection and diagnosis of pathologies. The objective is to replicate the diagnostic capabilities of a medical specialist based on patient data and where possible provide a more reliable diagnosis. In this endeavour, two trends have emerged: the first being pathology detection and the second being biological system modeling. The former has enjoyed cautionary clinician acceptance as pattern recognition is becoming common while the latter is less accepted due to the "black box" model which makes them difficult to be directly related to the pathology and thus remains a major issue to be addressed.

A general intelligent system incorporating CI consists of biosignal inputs, preprocessing, feature extraction, feature selection followed by pattern recognition (Figure 1). Biosignal inputs consist of measured signals from organs such as the heart or muscle. The signals are recorded, amplified, and filtered before feature extraction is performed. Proper feature extraction is essential for the later recognition stage as features should represent the unique pathological characteristics as much as possible. Feature selection is optional but functions to reduce redundant information which could confuse the detection process. The function of CI in this system lies in the final pattern recognition stage, where the task is to learn the implicit relationship between features and the respective pathologies given the finite set of data. It is important to note that the success of the system lies in its ability to perform well on new data and minimize the risk of misdiagnosis.

One of the more successful systems has been the diagnosis of cardiovascular diseases based on the QRS wave complex recorded from electrocardiograms (ECG) (Julian, Campbell, Cowan, & McLenachan, 2005). The average peak-to-peak voltage of the QRS complex is considerably larger than electrical activity from neighbouring organs making filtering simple and subsequent QRS detection algorithms sufficiently accurate (Friesen, Jannett, Jadallah, Yates, & Nagle, 1990). QRS information such as ST segment length (Edenbrandt, Devine, & Macfarlane, 1993), peak amplitudes, number of waveform turns, and peak to peak intervals (Kundu, Nasipuri, & Basu, 1998; Lin & Chang, 1989; Marques, Goncalves, Ferreira, & Abreu-Lima, 1994) have been effectively used as discriminating features. More advanced processing methods such as wavelet transforms, autoregression, and principal components analysis have also been applied to extract better features (Baxt,1993; Strauss, Steidl, & Jung, 2001). Classification of cardiovascular disorders using CI techniques, for example, ectopic beat detection (Chow, Moody, & Mark, 1992), arrhythmias (Coast, Stern, Cano, & Briller, 1990; Silipo, Zong, & Berthold, 1999), myocardial infarctions (heart attacks) (Fricker, 1997), and ischemia (Vladutu, Papadimithou, Mavroudi, & Bezerianos, 2001) has been sucessfully demonstrated. Tests performed on a variety of publicly available databases, for example, MIT-BIH[2] have yielded on average high accuracies with neural networks being the major applied technique (Table 1).

The classification of neuromuscular diseases has been more complex. Unlike the QRS waveform the motor unit action potential (MUAP), waveforms measured via electromyography (EMG) are composite waveforms derived from superposition of nerve signals emanating from adjacent muscle fibres (Emly, Gilmore, & Roy, 1992). Direct measurement of a single MUAP using needle EMG may not yield clear isolated waveforms either because motor units in the muscle fibers produce comparably smaller action potentials (-35mV to -70mV) making them easily corrupted by electrical noise. Even though muscle mechanics and action potentials are well understood, motor neuron firing is still thought to be random and nondeterministic. Research has examined the physical measurement of a single MUAP such as the number of turns and peak to peak amplitudes similar to QRS analysis (Park & Lee, 1998). This could be adequate provided a single clear MUAP waveform

Figure 1. General flowchart of intelligent detection biomedical system

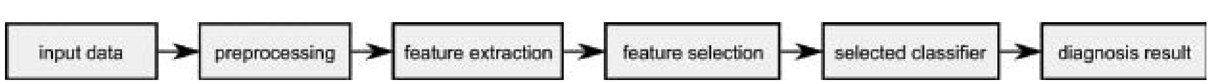


Table 1. Comparison of classification accuracies for ECG beat detectors using various classifier schemes

Classifiers	No.of Beats	Accuracy (%)
MLP1 (Hu, Thompkins, Urrusti, & Alfonso, 1993)	13	84.50
MLP2 (Izeboudjen & Farah, 1998)	12	90.00
SOM-LVD (Yu, Palreddy, & Thompkins, 1997)	4	92.20
MLP-LVQ (Oien, Bertelsen, Eftestol, & Husoy, 1996)	2	96.80
Hybrid Fuzz NN (Osowski & Tran, 2001)	7	96.06

could be obtained which is unfortunately rarely possible in practice. In this aspect, autoregressive (AR) models (Coatrieux, 1983; Pattichis & Elia, 1999) and independent component analysis (ICA) (Garcia, Martinez, Sornmo, Olmos, Mur, & Laguna, 2002) have been successfully employed. The former is based on a linear filter system and attempts to model the MUAP waveforms using a linear predictor model, while the latter performs MUAP decomposition but is highly sensitive to the final mixing matrix. The AR model performs better with higher model orders while ICA has been used in conjunction with template matching systems for the decomposition of MUAPTs (Garcia et al., 2002). Other methods such as the popular wavelet decomposition have also been applied, but it requires correct wavelet selection to extract meaningful information which is usually achieved through trial and error (Kumar, Pah, & Bradley, 2003). Much research is focused on finding more accurate signal processing techniques to interpret the MUAP trains (MUAPTs). Another issue is the type of features that should be extracted given that it may not be possible to recover the single MUAPs. In terms of classification, the literature has reported success in using neural networks (Pattichis, Schizas, & Middleton, 1995) while support vector machine (SVM) classifiers and hybrid models remain to be researched with few successful results. The emphasis for now is on using feature extraction methods to determine neuromuscular disorders for use in diagnosing diseases for example, myopathy and neuropathy from healthy subjects (Schizas, Pattichis, Schofield, Fawcett, & Middleton, 1990) (see Table 2).

In addition to neuromuscular diseases, EMG signals can be used with CI techniques in the development of prostheses and orthotics. Early myoelectric prostheses such as the Utah arm (Jacobsen, Knutti, Johnson, & Sears, 1982) possessed simple logic and few degrees of movement and depended on detecting a muscle posi-

Table 2. Comparison of classification accuracies of myopathy, motor neuron disease, and normal subjects. Note that the data set used is equivalent and the results are a comparison of the performance of these techniques and the corresponding features used.

Classifier Method	Feature Set	Accuracy (%)
BP-NN (Pattichis, Schizas, & Middleton, 1995)	MUAP time domain	80.00
NN-majority vote (Pattichis et al., 1995)	AR, cepstrum, time domain	80.00
Neurofuzzy hybrid (Christodoulou & Pattichis, 1999)	AR, cepstrum, time domain	88.58
Genetics Based Machine Learning (Pattichis & Schizas, 1996)	MUAP time domain	<95.00
Hybrid SOM-LVQ (Christodoulou & Pattichis, 1999)	MUAP time domain	97.61

tion after the user had flexed and held the position for some time. This approach was inadequate for situations requiring instant responses and a lot of time was spent on training the amputee to use the prosthesis. Recent research has concentrated on using CI to recognize limb positions (Karlik, Osman, & Alci, 2003; Katutoshi, Koji, & Takao, 1992; Peleg, Braiman, Yom-Tov, & Inbar, 2002) with very few studies of muscle activity when the limb is moving (Boca & Park, 1994). A more advanced prosthesis is the exoskeleton which provides mechanical strength either for everyday use or rehabilitative purposes. Kiguchi, Tanaka, and Fukuda (2004), for example, have developed an exoskeleton with three degrees of freedom (Kiguchi et al., 2004) and used a neurofuzzy system to detect the intended motion from EMG signals. More work, however, is required to improve the accuracy of translating and predicting the user's intention to move the exoskeleton, especially when detection errors can lead to injuries from unintended motion.

Electroencephalogram (EEG) signals from the brain provide important information regarding brain function and cognitive performance and can be detected based on their frequency bands, for example Alpha (8-12 Hz) and Beta (12-26 Hz) waves (Bylsma, Peyser, Folstein et al., 1994). Variations in these rhythmic sinusoidal signals can be used with CI techniques to detect various neurological disorders such as epilepsy (D'Alessandro, Esteller, Vachtsevanos et al., 2003), Alzheimer's (Cho, Kim, Park et al., 2003) and Huntington's (De Tommaso, Difruscoloand et al., 2003) diseases. Since these signals have small magnitudes, a major concern has been in noise reduction before CI detection. Another interesting area is the development of EEG-based communication links known as the brain controlled interfaces (BCI), which can restore functional activities in the disabled. CI techniques are proving very useful for classifying mental tasks from EEG signals which have subsequently improved the control of external devices and actuators (Deng & He, 2003; Felzer & Freisieben, 2003; Wang, Zhang, Li et al., 2004).

CI techniques have only recently been applied to research in human movement sciences and gait analysis upon the realization that locomotion is vital to our daily activities and our perceived quality of life (Whittle, 2002). Numerous gait variables such as joint angles (Oberg, Karsznia, & Oberg, 1994) and ground reaction forces (Begg, Sparrow, & Lythgo, 1998) can be recorded making feature selection an important aspect of gait data classification. CI has been recently used to study normal gait biomechanics (Barton & Lees, 1997; Barton, Lisboa, Lees, & Attfield, 2007), detect gait pathologies arising from lower extremity injuries and imbalance disorders due to cerebral diseases such as Parkinson's (Morris, Iansek, Matyas, & Summers, 1994) and Cerebral Palsy (Kamruzzaman & Begg, 2006). Falls in the elderly are another major health concern since they cause serious injuries and can be fatal, hence motivating the need for newer prediction and prevention techniques (National Safety Council, 2005). Recently CI techniques have been used to detect and predict elderly persons at risk of falling using gait variables such as the minimum toe clearance (Begg, Lai, Taylor, & Palaniswami, 2005; Khandoker, Lai, Begg, & Palaniswami, 2006). The uptake of CI approaches in gait analysis however has not kept pace with other biomedical areas and further research is required before these techniques become routine in clinical and rehabilitative settings.

WIRELESS SENSOR NETWORKS FOR HEALTH CARE

One emerging technology in health care is wireless sensor networks for patient monitoring systems aimed at reducing health care costs and addressing the shortage of medical staff in industrialized countries. The great advantage of a true wireless architecture is that there is little to no requirement for static infrastructure such as wires, cables, access points, and transmitting antennas to modems and routers. Wireless networks can be deployed using portable devices which communicate via radio frequency transmission with minimal infrastructure. We review briefly several existing research initiatives for sensor networks in health care.

Wireless Protocols and Standards

Several wireless technologies are currently in use such as wireless LAN (IEEE 802.11) and Bluetooth, which now come standard in personal computers, digital assistants (PDAs), and mobile phones. Wireless LAN has been used in hospitals for Internet access and to transfer patient data between departments, and now enjoys widespread acceptance due to its extensive coverage and ease of use. Unfortunately, wireless LAN requires high power consumption, large packet overheads, and requires static infrastructure support such as access points. Bluetooth on the other hand was designed as a short range protocol for communication between a small group of devices and is ideal for office spaces where data can be transmitted to devices within a 10m to 50m radius (Haartsen, 2000). It was envisaged that Bluetooth would be used in the deployment of Wireless Personal Area Networks (WPANs) in workplaces but was later found to be susceptible to degradation due to noise and multipath effects which reduced its effective range. This protocol nevertheless features prominently in mobile handsets and is sufficient for the local exchange of music, pictures, and data files. Both protocols look unlikely to be suitable for the design of wireless health care monitoring systems incorporating small sensor nodes due to a different set of requirements.

The key difference between wireless computing networks and sensor networks is that small low computing powered devices (sensors) are equipped with wireless capabilities to form a network. These devices are intended to collect data from the body in our case, and transmit it to a processing centre (such as a PC) usually called a mother node. Because the sensor nodes are small they do not possess substantial on board energy for transmission which restricts them to smaller volumes of data (low data rates) over shorter distances. Newer protocols, such as the IEEE 802.15.4 or Zigbee, have been proposed to address these design requirements. Zigbee was conceived in 1998 and completed in 2003 and is primarily designed for low powered devices over short ranges. This type of protocol is well suited to wireless monitoring systems for health care (Golmie, Cypher, & Rebala, 2004, 2005), for example, a wireless ECG monitor will allow a patient to move around the hospital while still being monitored. Rapid action can be taken if complications arise while providing more mobility and privacy for the patient. These networks are also known as Body Area Networks (BANs) comprising unobtrusive devices to monitor health signs such as heart rate or blood pressure (Guang-Zhong, 2006; Jovanov, Milenkovic, & Otto, 2005; Jovanov, Milenkovic, Otto, & Groen, 2005).

Another emerging technology is Radio Frequency Identification (RFID), a labeling technology in retail which has been increasingly deployed (Roy, 2006) for tracking. When combined with databases and globally linked systems, RFID can be used pinpoint patients or pharmaceuticals as they move in the hospital. These features are the driving force behind continued research into RFID technology for health care since a hospital could, for example, use RFID to tag blood samples to the patient thereby eliminating the need for paper records.

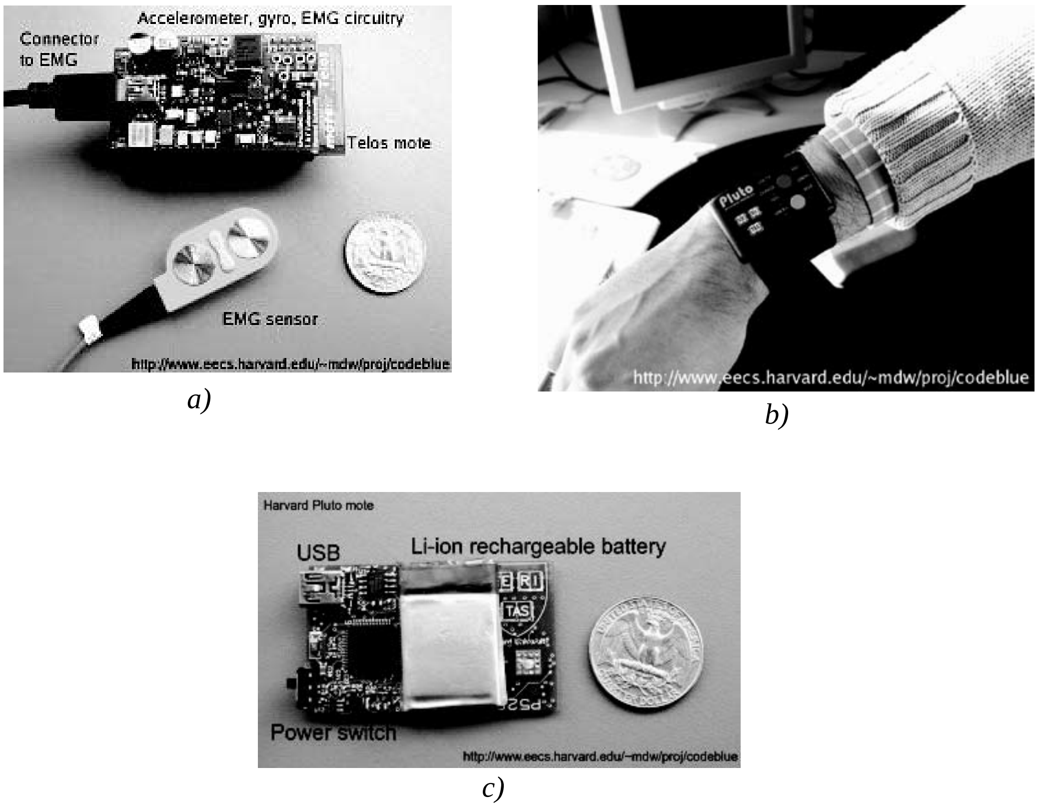
CodeBlue: Wireless Sensor Networks for Medical Care

Harvard University has initiated a wireless network project for health care called Codeblue (http://www.eecs.harvard.edu/ mdw/proj/codeblue/) in collaboration with hospitals (such as the Boston Medical Center) with support from national institutions such as the National Science Foundation (NSF), National Institutes of Health (NIH) and also companies such as Microsoft. The aim of the project, which began in 2003, is to explore applications of wireless sensor network technology to a range of medical applications, including prehospital and in-hospital emergency care, disaster response, and stroke patient rehabilitation.

Several small sensor devices with wireless capabilities such as a wireless pulse oximeter based on the Mica and Telos sensor motes from Crossbow and the University of Berkeley respectively have been developed. Other devices include a modified Telos mote known as the Pluto mote equipped with accelerometers, gyroscopes, and EMG circuitry (Figure 1). The mote is incorporated in a wrist-worn device and monitors vital health statistics such as heart rate. A Codeblue software platform provides routing, naming, discovery, and security for wireless medical sensors, PDAs, PCs, and other devices for health care monitoring (Lorincz, Malan, Fulford-Jones et al., 2004). A subsystem for this platform is MoteTrack, used for tracking the location of individual patient devices both indoor and outdoor using radio signals from beacons fixed in the surroundings. The group is focused on major issues in the integration of medical sensors with low-power wireless networks such as wireless routing, security, 3D location tracking, and adaptive resource management.

Other groups adopting Codeblue software include, for example, the Motion Analysis Laboratory in the Spaulding Rehabilitation Hospital[3] (and the AID-N project at the Johns Hopkins Applied Physics Laboratory,[4] which is investigating a range of technologies for disaster response. The AID-N wireless sensors are deployed on devices such as an electronic “triage tag” with pulse oximeter, LCD display, and LEDs for

Figure 2. (a) The combined accelerometer, gyroscope, and electromyogram (EMG) sensors for stroke patient monitoring, developed by the Codeblue project. (b) The Pluto mote can be mounted in a case which makes it easy for wearable wrist devices. (c) The Pluto mote designed at Havard. The architecture is based on the design of the Telos mote from the University of Berkeley and uses rechargeable Lithium ion batteries (http://www.eecs.harvard.edu/~mdw/proj/codeblue/ accessed 4/15/2007).



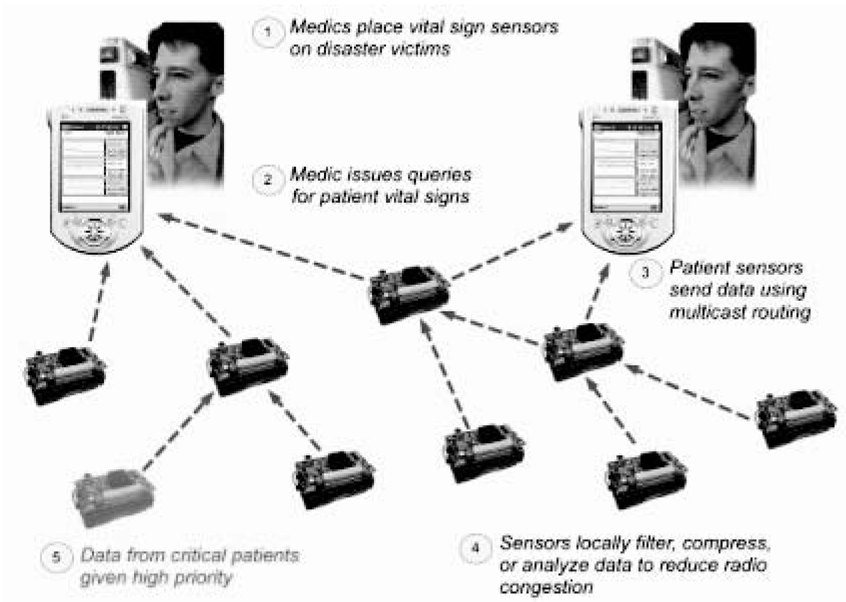
indicating a patient's status. There is also a packaged version of the two lead ECG mote developed by the Harvard group and also a wireless blood pressure cuff. Other collaborations include a wide-area event delivery infrastructure for medical care, with Harvard University, and a project integrating Vital Dust sensors into a PDA-based patient record database (iRevive). The iRevive system provides real-time monitoring of vital signs by emergency department personnel and clinicians through a variety of interfaces such as PDAs or Web-based PCs.

Mobihealth: Body Area Networks (BANs) using Wireless Mobile Phone Technology

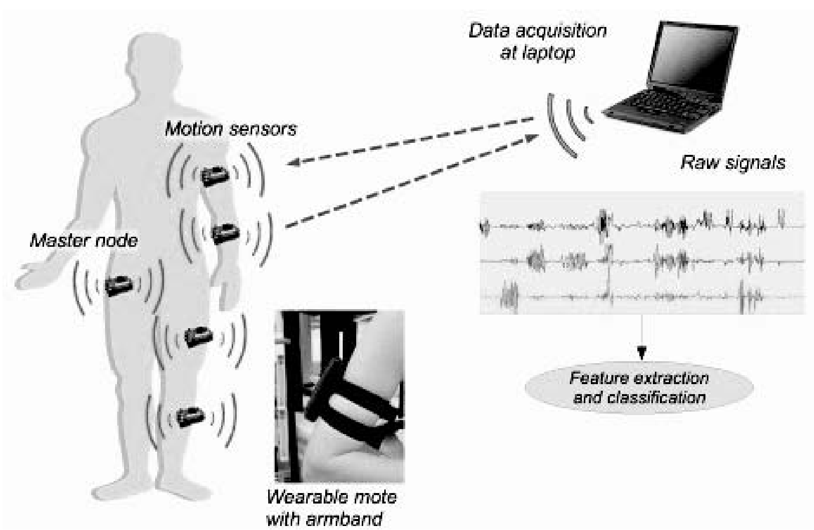
Mobihealth (2002) (http://www.mobihealth.org/) is a collaborative project based on a European initiative to create a generic platform for home health care. The platform defines a body area network (BAN) with low bandwidth GPRS/UMTS wireless protocols for communication. The BAN devices can be worn by patients and serve to monitor vital health statistics, being both low-powered and customizable. The technology is directed at providing new services to the remote monitoring of chronically ill patients (in hospital and at home), remote assistance in emergency and disaster situations, and the remote management of clinical trials (research augmentation). It is hoped that these developments will reduce health care costs while maximizing response times. Project designers have focused on developing a generic platform which could incorporate all these services.

From a CI perspective, Mobihealth presents an exciting new possibility to design a platform supporting intelligent plug-and-play devices. One example is to incorporate CI for early warning detection of high risk patients, for example cardiac health. This presents an

Figure 3. (a) The Codeblue architecture consisting of Telos or Pluto motes used for emergency response, (b) limb motion monitoring for rehabilitation in stroke patient. Photo courtesy of Codeblue homepage (http://www.eecs.harvard.edu/~mdw/proj/codeblue/ Accessed 4/15/2007).



a)



b)

opportunity for the application of CI techniques with BAN technology not only for abnormality detection but also for onset prediction of a serious pathology. Detection and prediction models can be placed on board the sensor devices to increase their advantages allowing immediate contact of emergency personnel and overall improved response times.

Other Health Care Sensor Research Initiatives

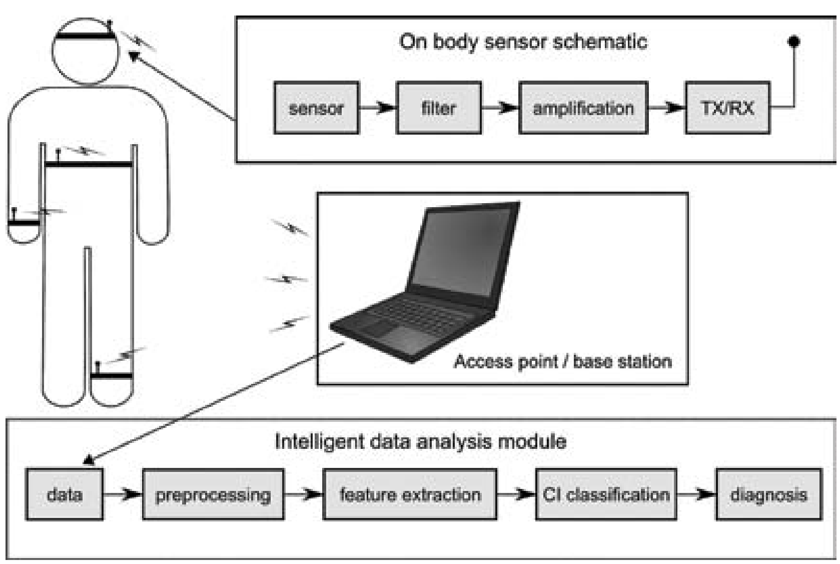
Several other projects have already commenced on using wireless technologies in health care, ranging from sensor deployment to the underlying transmission protocols undertaken by universities in collaboration with hospitals. The UPDATA project (http://www.sscnet.ucla.edu/issr/da/PDA/) was initiated by the biomedical library at the University of California Los Angeles (UCLA), with the objective of incorporating statistical analysis in medical data via a PDA. The project is attempting to create a central server with customizable searches, calculation tools (such as ratios and dosages), and a database of links to major health care sources all of which can be accessed by researchers using PDAs. Another project (Jovanov, Milenkovic, & Otto, 2005; Jovanov, Milenkovic, Otto et al., 2005) has investigated wireless sensor technology for ambulatory and implantable psycho-physiological applications. These devices are capable of monitoring the functioning of the heart, long time motion of prosthetic joints, and for monitoring other vital signs. Their main focus is wearable and implantable (or subcutaneous) biosensors, biomaterial technologies, modelling, and wireless communication. Further scope exists for deploying CI as backend software either for assisting diagnosis or for data interpretation.

Research initiatives involving personalized and non-hospital based care are also receiving significant attention. It has been estimated that two thirds of deaths due to cardiac disease occur outside hospital environments (Rubel, Fayn, Simon-Chautemps et al., 2004). This suggests the need for personalized diagnostic devices that can monitor ECG waveforms in the community and warn of any forthcoming cardiac events. Such a system will require intelligent information processing to analyse patient vital signs and alert the care providers. Rubel et al. (2004) reported the development of an intelligent Personal ECG Monitor (PEM) for early detection of cardiac events. Their system is equipped with in-built advanced decision making and can generate alarm levels and messages for care providers. This type of device demonstrates how effectively pervasive computing and CI-based intelligent information processing, combined with personalized and wearable ubiquitous devices, could improve health care.

INTELLIGENT SENSOR NETWORK SYSTEMS FOR HEALTH CARE MONITORING

In this section, we discuss the properties of a health care system incorporating computational intelligence and sensor network technology in a single paradigm. A general system (Figure 4) consists of the sensor network, first deployed either by attaching sensor nodes to a patient or placing them in the patient area. Once the network is setup, the actual measurements can be made, collected, and transmitted to a base node for further processing. Measurement data is generally corrupted by noise and must be filtered and preprocessed using signal processing techniques. The processed data can then be analysed with a selection of CI techniques and the appropriate decisions made based on the system outputs.

Figure 4. General schematic of an intelligent sensing system for patient monitoring



The advantage of sensor network monitoring compared to traditional bed care monitoring is that measurements can be continuously made and the patient need not be confined to the bed. This is advantageous for diseases such as diabetes (Klonoff, 2005) where patients still value their mobility. A recent example application successfully incorporating intelligent sensor network monitoring is the determination of levodopa-induced dyskinesias (LID) (Keijsers, Horstink, & Gielen, 2003). LIDs are characterized by involuntary muscle movements caused by levodopa medications used to treat Parkinson's disease. The long term side effect is involuntary muscle movements ranging from unnoticeable to severe. These movements are measured using the Abnormal Involuntary Movement Scale (AIMS) with score ranges from 0 to 4, with 4 being the most severe. Knowledge of the amount of LIDs is useful for estimating them accurately to improve the daily life of the patient. We now review the components in this system.

Sensor Measurements

The measurement system consists of six accelerometers attached to the patient's body: one on each foot, two on the shoulders, one on the chest, and one on the arm with the most severe LID. The small sensors can be quickly attached and do not interfere with patient mobility. The nodes serve to measure the linear acceleration which can be processed to obtain velocity and displacement measurements. Since the main symptom of LIDs is uncontrollable muscle tremors, the acceleration data is converted in to the amount and direction of low frequency vibrations (3 Hz or less). In general, signal processing algorithms such as wavelets (Adeli, Ghosh-Dastidar, & Dadmehr, 2007) and principal component analysis (Pawar, Anantakrishnan, Chaudhuri, & Duttagupta, 2007) could also be used to extract useful information giving a wider scope for this and other applications. Keijsers et al. (2003) derived 92 features in total, ranging from percentage of time the patient was sitting to various cross-correlations between different sensors.

Data Analysis

The system is fully off-line; data is logged and then processed on a separate PC similar to other processing techniques such as electrocardiogram (Paoletti & Marchesi, 2006), patient sleep state (Held et al., 2006), metabolism (Davies, Batholomeousz, & Andrade, 2003), and urea (Paoletti & Marchesi, 2006). This allows for a computer implementation of neural networks for supervised learning, specifically the multilayer perceptron (MLP) trained with the standard backpropagation algorithm. The size of the network was determined using forward selection as suggested by Laar, Heskens, and Gielen (1999). This classifier may be replaced with other computational intelligence methods such as the self-organizing map (Bonato, Mork, Sherrill, & Westgaard, 2003), support vector machines (Ubeyli, 2007), and spiking neurons (Rom et al., 2007).

Traditional LID measurements were previously done by asking the patient to remain still so that the measured movements could be assumed to originate from involuntary actions. Detection of involuntary motions becomes more complex when a patient is allowed every day motions such as cleaning or making coffee. The reliable detection of involuntary movements from voluntary ones would require more complex algorithms such as artificial neural networks. In addition, changes to the degree of these movements could depend on several other biomechanical factors which could be extremely difficult to explicitly model. Computational intelligence affords us a method to approximate these changes by learning the relationships between the effects and the system outcomes hence making detection easier.

System Performance

The performance of Keijsers et al. (2003) system was compared against two physicians who assessed the patient's AIMS score (0 to 4) from video tapes. Their classifications were considered as the true detection and the neural network was trained and tested against this. Results indicated that the system could consistently and reliably estimate LID value from sensor measurements. The predicted AIMS scores were never more than 1 off the physician scores and in 98% of the cases were correct. As the tests were cross-validated, they showed that the system could generalize to unseen samples of the same patient as well as to new patients which were encouraging because it demonstrated that sensor data could be reliably integrated with computational intelligence to design more intelligent monitoring systems.

CONCLUSION

In this article, we reviewed the applications of two emerging technologies in health care, namely computational intelligence and sensor networks. Both these technologies offer cheap and efficient alternatives to the current shortages in medical personnel faced by today's global health care systems. The combination of these technologies has been applied to the development of intelligent detection and patient monitoring systems, thereby reducing the patient attention demands on clinicians. Further research is required to develop more accurate and reliable systems which can provide similar or better qualities of care than current health care systems. The future looks promising as work moves to make both technologies ubiquitous in health care systems.

REFERENCES

Abe, S. (1997). *Neural networks and fuzzy systems: Theory and applications*. Boston: Kluwer Academic.

Adeli, H., Ghosh-Dastidar, S., & Dadmehr, N. (2007, February). A wavelet-chaos methodology for analysis of EEGs and EEG subbands to detect seizure and epilepsy. *IEEE Transactions on Biomedical Engineering, 54*(2), 205-211.

Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., & Cayirci, E. (2002). A survey on sensor networks. *IEEE Communications Magazine, 1,* 102-114.

Barton, G., Lisboa, P., Lees, A., & Attfield, S. (2007). Gait quality assessment using self-organising artificial neural networks. *Gait and Posture, 25,* 374-379.

Barton, J., & Lees, A. (1997). An application of neural networks for distinguishing gait patterns on the basis of hip-knee joint angle diagrams. *Gait and Posture, 5,* 28-33.

Baxt, W. (1993). Classification of cardiac arrhythmias based on pricipal component analysis and feedforward neural networks. *Computers in Cardiology,* 341-344.

Begg, R., Lai, D., Taylor, S., & Palaniswami, M. (2005). SVM models in the diagnosis of balance impairments. In *3rd International Conference on Intelligent Sensing and Information Processing,* Banglore, India (pp. 248-253).

Begg, R., Sparrow, W., & Lythgo, N. (1998). Time-domain analysis of footground reaction forces in negotiating obstacles. *Gait and Posture, 7,* 99-109.

Boca, D., & Park, D. (1994). Myoelectric signal recognition using fuzzy clustering and artificial neural networks in real time. *IEEE International Conference on Neural Networks, 1,* 3098-3103.

Bonato, P., Mork, P.J., Sherrill, D.M., & Westgaard, R.H. (2003, May-June). Data mining of motor patterns recorded with wearable technology. *IEEE Engineering in Medicine and Biology Magazine, 22*(3), 110-119.

Bylsma, F., Peyser, C., Folstein, S. et al. (1994). EEG power spectra in Huntington's disease: Clinical and neuropsychological correlates. *Neuropsychologia, 32*(2), 137-150.

Cho, S., Kim, B., Park, E. et al. (2003). Automatic recognition of Alzheimer's disease with single channel eeg recording. In *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (Vol. 3, pp. 2655-2658).

Chong, C.-Y., & Kumar, S. (2003). Sensor networks: Evolution, opportunities, and challenges. In *Proceedings of the IEEE, 91*(8), 1247-1256.

Christodoulou, C., & Pattichis, C. (1999). Unsupervised pattern recognition for the classification of EMG signals. *IEEE Transactions on Biomedical Engineering 46*(2), 169-178.

Coast, D., Stern, R., Cano, G., & Briller, S. (1990). An approach to cardiac arrhythmia analysis using hidden markov models. *IEEE Transactions on Biomedical Engineering, 37*(9), 826-836.

Coatrieux, J. (1983). Interference electromyogram processing- Part II: Experimental and simulated EMG AR modeling. *Electromyography Clinical Neurophysiology, 23,* 481-490.

D'Alessandro, M., Esteller, R., Vachtsevanos, G. et al. (2003). Epileptic seizure prediction using hybrid feature selection over multiple intracranial EEG electrode contacts: A report of four patients. *IEEE Transactions on Biomedical Engineering, 50,* 603 615.

Davies, R., Batholomeousz, D. A., & Andrade, J. (2003, January/February). Personal sensors for the diagnosis and management of metabolic disorders. *IEEE Engineering in Medicine and Biology Magazine, 22*(1), 32-42.

Deng, J., & He, B. (2003). Classification of imaginary tasks from three channels of EEG by using an artificial neural network. In *Proceedings of the 25th Annual International Conference of the IEEE* (Vol. 2, pp. 2289-2291).

De Tommaso, F.D.C., M.M., Difruscoloand, O. et.al. (2003). Detection of subclinical brain electrical activity changes in Huntington's disease using artificial neural networks. *Clinical Neurophysiology, 114*(7), 1237-1245.

Edenbrandt, L., Devine, B., & Macfarlane, P. (1993). Neural networks for classification of ECG ST-T segments. *Journal of Electrocardiology, 26*(3), 239-40.

Emly, M., Gilmore, L., & Roy, S. (1992). Electromyography. *IEEE Potentials, 11*(2), 25-28.

Felzer, T., & Freisieben, B. (2003). Analyzing EEG signals using the probability estimating guarded neural classifier. *IEEE Transactions on Neural Systems and Rehabilitation engineering, 11,* 361- 371.

Fricker, J. (1997). Artificial neural networks improve diagnosis of acute myocardial infarction. *Lancet, 350,* 935.

Friesen, G., Jannett, T., Jadallah, M., Yates, S., Quint, S.L., & Nagle, H. (1990). A comparison of the noise sensitivity of nine QRS detection algorithms. *IEEE Transactions on Biomedical Engineering, 37*(1), 85-98

Garcia, J., Martinez, I., Sornmo, L., Olmos, S., Mur, A., & Laguna, P. (2002). Remote processing server for ECG-based clinical diagnosis support. *IEEE Transactions on Information Technology in Biomedicine, 6*(4), 277-284.

Golmie, N., Cypher, D., & Rebala, O. (2004). Performance evaluation of low rate wpans for sensors and medical applications. In *Military Communications Conference (MILCOM).*

Golmie, N., Cypher, D., & Rebala, O. (2005). Performance analysis of low rate wireless technologies for medical applications. *Computer Communications, 28*(10), 1255-1275.

Guang-Zhong, Y. (2006). *Body sensor networks.* New York: Springer.

Haartsen, J.C. (2000). The Bluetooth radio systems. *IEEE Personal Communications, 7*(1), 28-36.

Haykin, S. (1994). *Neural networks: A comprehensive foundation*. Prentice Hall.

Held, C.M., Heiss, J.E., Est´evez, P.A., Perez, C.A., Garrido, M., Algarn, C. et al. (2006, October). Extracting fuzzy rules from polysomnographic recordings for infant sleep classification. *IEEE Transactions on Biomedical Engineering, 53*(10), 1954-1962.

Hu, Y., Tompkins, W., Urrusti, J., & Afonso. V. (1993). Applications of artificial neural networks for ECG signal detection and classification. *Journal of Electrocardiolgy, 26,* 66-73.

Izeboudjen, N., & Farah, A. (1998). A new neural network system for arrhythmia's classification. In *Proceedings of the Neural Network Conference* (Vol. 1, pp. 208-216).

Jacobsen, S., Knutti, D., Johnson, R., & Sears, H. (1982). Development of the Utah artificial arm. *IEEE Transactions on Biomedical Engineering, 29,* 249-269.

Jovanov, E., Milenkovic, A., & Otto, C. (2005). A WBAN system for ambulatory monitoring of physical activity and health status: Applications and challenges. In *Proceedings of the 27th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)* (Vol. 1(1), pp. 3810-3813).

Jovanov, E., Milenkovic, A., Otto, C., & Groen, P. de. (2005). A wireless body area network of intelligent motion sensors for computer assisted physical rehabilitation. *Journal of Neuroengineering and Rehabilitation, 2*(6).

Julian, D.G., Campbell Cowan, J., & McLenachan, M.J. (2005). *Cardiology* (8^{th} ed.). Edinburgh, NY: Saunders.

Kamruzzaman, J., & Begg, R. (2006). Support vector machines and other pattern recognition approaches to diagnosis of cerebral palsy gait. *IEEE Transactions on Biomedical Engineering, 52,* 828-838.

Karlik, B., Osman, T., & Alci, M. (2003). A fuzzy clustering neural network architecture for multifunction upper-limb prosthesis. *IEEE Transactions on Biomedical Engineering, 50*(11), 1255- 1261.

Katutoshi, K., Koji, O., & Takao, T. (1992). A discrimination system using neural network for EMG-controlled prosthesis. In *Proceedings of the IEEE International Workshop on Robot Human Communication* (Vol. 1, pp. 63-68).

Keijsers, N.L., Horstink, M.W., & Gielen, S.C. (2003). Automatic assessment of levodopa-induced dyskinesias in daily life by neural networks. *Movement Disorders, 18*(1), 70-80.

Khandoker, A., Lai, D., Begg, R., & Palaniswami, M. (2006). A wavelet based approach for screening falls risk in the elderly using support vector machines. In *IEEE Proceedings of Fourth International Conference on Intelligent Sensing and Information Processing (ICISIP)* (Vol. 1, pp. 78-82).

Kiguchi, K., Tanaka, T., & Fukuda, T. (2004). Neuro-fuzzy control of a robotic exoskeleton with EMG signals. *IEEE Transactions on Fuzzy Systems, 12*(4), 481-490.

Klonoff, D.C. (2005, May). Continuous glucose monitoring. *Diabetes Care, 28*(5), 1231-1239.

Kumar, D., Pah, N., & Bradley, A. (2003). Wavelet analysis of surface electromyography. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 11*(4), 400- 406.

Kundu, M., Nasipuri, M., & Basu, D. (1998). A knowledge-based approach to ECG interpretation using fuzzy logic. *IEEE Engineering in Medicine and Biology Magazine, 28*(2), 237-243.

Laar, P., Heskens, T.M., & Gielen, C.C.A.M. (1999). Partial retraining: A new approach to input relevance determination. *International Journal of Neural Systems, 9*, 75-85.

Lin, K.-P., & Chang, W. (1989). QRS feature extraction using linear prediction. *IEEE Transactions on Biomedical Engineering, 36*(10), 1050-1055.

Lorincz, K., Malan, D., Fulford-Jones, T. et al. (2004). Sensor networks for emergency response: Challenges and opportunities [Special issue on pervasive computing for first response]. *IEEE Pervasive Computing, 4*(1), 16-23.

Marques, J., Goncalves, A., Ferreira, F., & Abreu-Lima, C. (1994). Comparison of artificial neural network based ECG classifiers using different feature types. *Computers in Cardiology, 1*, 545-547.

Morris, M., Iansek, R., Matyas, T., & Summers, J. (1994). The pathogenesis of gait Hypokinesia in Parkinson's disease. *Brain, 117*(5), 1169-1181.

National Safety Council. (2005). *Falls among older adults: A national problem*. White Paper for White House Conference on Aging.

Oberg, T., Karsznia, A., & Oberg, K. (1994). Joint angle parameters in gait: Reference data for normal subjects, 10-79 years of age. *Journal of Rehabilitation Research and Development, 31*, 199-213.

Oien, G., Bertelsen, N., Eftestol, T., & Husoy, J. (1996). ECG rhythm classification using artificial neural networks. In *IEEE Digital Signal Processing Workshop* (Vol. 1, pp. 514-517).

Osowski, S., & Tran, H. (2001). ECG beat recognition using fuzzy hybrid neural network. *IEEE Transactions on Biomedical Engineering, 48(11)*, 1265-1271.

Paoletti, M., & Marchesi, C. (2006). Discovering dangerous patterns in longterm ambulatory ECG recordings using a fast qrs detection algorithm and explorative data analysis. *Computer Methods and Programs in Biomedicine, 82*, 20-30.

Park, S.-H., & Lee, S.-P. (1998). EMG pattern recognition based on artificial intelligence techniques. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 6*(4), 400-405.

Pattichis, C., & Elia, A. (1999). Autoregressive and cepstral analyses of motor unit action potentials. *Elsevier Journal of Medical Engineering and Physics, 21*, 405-419.

Pattichis, C., & Schizas, C. (1996). Genetics-based machine learning for the assessment of certain neuromuscular disorders. *IEEE Transactions on Neural Networks, 7*(2), 427-439.

Pattichis, C., Schizas, C., & Middleton, L. (1995). Neural network models in EMG diagnosis. *IEEE Transactions on Biomedical Engineering, 42*(5), 486-496.

Pawar, T., Anantakrishnan, N.S., Chaudhuri, S., & Duttagupta, S.P. (2007, June). Transition detection in body movement activities for wearable ECG. *IEEE Transactions on Biomedical Engineering, 54*(6), 1149-1152.

Peleg, D., Braiman, E., Yom-Tov, E., & Inbar, G. (2002). Classification of finger activation for use in a robotic prosthesis arm. *IEEE Transactions on Rehabilitation Engineering, 10*(4), 290-293.

Rom, R., Erel, J., Glikson, M., Lieberman, R.A., Rosenblum, K., Binah, O. et al. (2007, March). Adaptive cardiac resynchronization therapy device based on spiking neurons architecture and reinforcement learning scheme. *IEEE Transactions on Neural Networks, 18*(2), 542-550.

Roy, W. (2006). An introduction to RFID technology. *IEEE Pervasive Computing, 5*(1), 25-33.

Rubel, P., Fayn, J., Simon-Chautemps, L. et al. (2004). New paradigms in telemedicine: Ambient intelligence, wearable, pervasive and personalized. In A. Lymberis & D. de Rossi (Eds.), *Wearable e-health systems for personalised health management: State of the art and future challenges*. Amsterdam: IOS Press.

Schizas, C., Pattichis, C., Schofield, I., Fawcett, P., & Middleton, L. (1990). Artificial neural nets in computer-aided macro motor unit potential classification. *IEEE Engineering in Medicine and Biology Magazine, 9*(3), 31-38.

Schlkopf, B., Burges, C., & Smola, A. (1999). *Advances in kernel methods: Support vector learning.* Cambridge, MA: MIT Press.

Silipo, R., Zong, W., & Berthold, M. (1999). ECG feature relevance in a fuzzy arrhythmia classifier. *Computers in Cardiology, 1*, 679-682.

Strauss, D., Steidl, G., & Jung, J. (2001). Arrhythmia detection using signal adapted wavelet preprocessing for support vector machines. *Computers in Cardiology,* 497-500.

Übeyli, E.D. (2007). Combining eigenvector methods and support vector machines for detecting variability of Doppler ultrasound signals. *Computer Methods and Programs in Biomedicine, 86,* 181-190.

Vapnik, V.N. (2000). *The nature of statistical learning theory* (2nd ed.). New York: Springer.

Vladutu, L., Papadimithou, S., Mavroudi, S., & Bezerianos, A. (2001). Ischemia detection using supervised learning for hierarchical neural networks based on kohonen-maps. In *Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (Vol. 2, pp. 1688-1691).

Wang, Y., Zhang, Z., Li, Y. et al. (2004). BCI competition 2003-data set iv: An algorithm based on CSSD and FDA for classifying single-trial EEG. *IEEE Transactions on Biomedical Engineering, 51*, 1081 1086.

Whittle, M. (2002). *Gait analysis: An introduction* (3rd ed.). Oxford: Butterworth Heinemann.

Yu, H.H., Palreddy, S., & Tompkins, W. (1997). A patient-adaptable ECG beat classifier using a mixture of experts approach. *IEEE Transactions on Biomedical Engineering, 44*(9), 891-900.

KEY TERMS

Classification: Assigning a class designation to a set of subject attributes.

Computational Intelligence: Mathematical formulations which simulate human learning and are implemented on computers.

Detection: Becoming aware of a present condition.

Diagnosis: Reaching a conclusion on a certain disorder based on observation and physical measurements.

Intelligent Monitoring Systems: Systems designed to monitor patient health signs and are capable to act on them.

Prognosis: The severity and future outlook of a disease and its progression.

Rehabilitation: The recovery process from a disorder after medical treatment.

Sensor Networks: A network of sensors which are each equipped with radio transceivers.

ENDNOTES

1 www.who.int/whf/2006/whr06 en.pdf
2 Online: http://ECG.mit.edu
3 http://spauldingrehab.org/
4 http://www.aid-n.org/

Computational Methods for the Early Detection of Diabetes

Herbert F. Jelinek
Charles Sturt University, Australia

David J. Cornforth
University of New South Wales, Australia

INTRODUCTION

The incidence of diabetes is increasing, and is expected to exceed one million people in Australia by the year 2010. Diabetes is currently diagnosed after the onset of symptoms. At this stage, detection of complications is a key intervention point in reducing the associated personal and community burden. These include eye, heart, kidney, and foot disease, which in many instances can be treated with good outcomes, provided the disease process is recognized early. In Australia, both national and state governments acknowledge the disadvantage faced by rural people in availing themselves of all aspects of diabetes management, from screening to regular assessment, education and health care (Colagiuri, Colagiuri, & Ward 1998). Therefore, the challenge that is the focus of this article is the early detection of diabetes complications associated with vision and cardiac function, with the eventual aim of providing a screening service that can be used in a rural or regional environment.

BACKGROUND

Complications associated with diabetes often remain undetected for quite some time, especially in rural regions, but are often what alerts health care providers to the presence of diabetes. Vision loss, cardiovascular disease, and foot amputations are the most often occurring outcomes of diabetes. Early detection of blood vessel changes associated with retinopathy, cardiovascular disease, and foot complications is important for timely treatment, to prevent or delay the occurrence of these complications and improve the quality of life for individuals with diabetes. Regular health screening that includes eye and cardiac function assessment has the potential to reduce the high costs on the health care system associated with diabetes, and to reduce the disparity in health care between rural and urban communities by providing timely feedback to those at risk.

Currently, regular screening provides an opportunity to discover people with unidentified diabetes and with subclinical eye and heart complications. Extrapolation from data obtained from eye health initiatives indicates that interventions that relieve or prevent complications associated with diabetes are highly cost effective (Lee, Lee, Kingsley, Wang, Russell, Klein, & Warn, 2001).

Universal screening is both feasible and cost effective. It delays complications, and younger patients have a longer time to benefit from early identification. In spite of this, diabetes screening is far from universal, especially among rural and regional populations. However, recent reductions in the cost of health technology and an increase in the availability of computing equipment, coupled with advanced data processing techniques, offer new opportunities. These include the ability to detect complications earlier, make screening faster, more accurate and cheaper, and to remove access barriers faced by those in rural and regional communities. This article focuses on two aspects of screening that could benefit from automation using computational techniques: diabetic retinopathy (DR), and cardiovascular disease.

DR is a condition in which the retina has suffered damage due to the complications of diabetes. DR could be significantly reduced by simplifying the procedure used to identify the condition and ensuring that early eye examinations become routine for diabetic patients. Damage to the eye associated with DR is the commonest cause of blindness in the working age population in developed countries. Yet, 98% of people with vision loss can be treated effectively provided the pathology is detected early (Lee et al., 2001). Automated assessment of retinal health allows community screening programs to be offered and operated effectively, utilizing community health professionals.

Cardiovascular disease includes a range of diseases, of which coronary artery disease (CAD) is the single most common cause of death in the western world. Cardiovascular complication is associated with diabetes, and on its own is the most common cause of death (Mathers & Penm, 1999). Cardiovascular disease affecting the larger blood vessels is a major complication of diabetes. People with diabetes are two to four times more likely to develop cardiovascular disease, and their prognosis is not as good as those without the disease. Cardiac autonomic neuropathy (CAN) is part of the cardiovascular disease spectrum, in part caused by CAD, which leads to reduced oxygenation of the nerves regulating the cardiac rhythm. The resulting morbidity and mortality can be reduced with regular screening and early recognition of nervous system damage affecting cardiac function in people with diabetes, which allow better treatment intervention.

Electrocardiology is still regarded as the most commonly used procedure for the assessment of cardiac function and identification of heart disease, and is utilized by numerous health care practitioners. The 12-lead electrocardiograph (ECG) is the kind of equipment most often used in regular general practice. In community health care, outreach nurses do not have the in-depth training to determine risk of a potentially dangerous cardiovascular condition, based on signs and symptoms, and are most often familiar with three-lead recordings. ECG interpretation is usually carried out by a specialist.

Screening for both cardiovascular disease and DR in rural areas is difficult because of health care barriers such as geographical isolation, the cost of visits to specialists, and the lack of cardiologists and ophthalmologists.

COMPUTATIONAL METHODS FOR THE EARLY DETECTION OF DIABETES IN A RURAL CONTEXT

Advances in computational methods that may be of use in rural settings involve two main areas: DR, and heart rate variability (HRV). Both of these rely on the fact that diabetes affects the cardiovascular system. Both aim to provide methods that can be applied easily and cheaply in a rural setting.

Analysis of Diabetic Retinopathy

The assessment of retinopathy has been enhanced by the ability of regional clinics to make quality photographs of the retina that can be then sent to remote ophthalmologists for assessment. This has recently been expedited using digital photography and improvements in digital communication. The best quality images are obtained using fluorescein photography, where the contrast is enhanced by using a dye injected into the bloodstream. This, however, is invasive, and much research has been conducted into processing of nonfluorescein color images using computational techniques.

The scope of DR covers a wide range of pathologies that are amenable to automated identification including:

- Microaneurysms (Jelinek, Cree, Worsely, Luckie, & Nixon, 2006; Cree, Olson, McHardy, Forrester, & Sharp, 1996).
- Hard exudates and cotton wool spots (Dua, Kandiraju, & Thompson, 2005).
- Haemorrhages (Singalavanija, Supokavej, Bamroongsuk, Sinthanayothin, Phoojaruenchanachai, & Kongbunkiat, 2006).
- New vessel growth (Soares, Leandro, Cesar, Jelinek, & Cree, 2006).
- Arteriolar narrowing and venous dilatation (Wong, Shankar, Klein, Klein, & Hubbard, 2005).

Automating the assessment of changes in the retina can be carried out using mathematical techniques. These techniques can be applied to the estimation of blood vessel diameter changes and quantify the appearance of microaneurysms and haemorrhages in the early stages of retinopathy, as well as new vessel growth in the advanced stages of DR. Of interest has been the correlation between blood vessel diameter changes in the eye that is not only associated with diabetes, but is also an early indicator for cardiovascular disease and stroke (Wong et al., 2005). For the automated assessment of most of these, the optic disk needs to be identified. This facilitates, for example, macula location, false positive localization of hard exudates, and vessel tracking.

Retinal Image Processing

Several computational operators have been utilized for retinal image processing. Morphological filtering simplifies a segmented image to facilitate the search for objects of interest. This is done by smoothing out object outlines, filling small holes, eliminating small projections, and other similar techniques.

Many image processing techniques use frequency filtering, based on the Fourier Transform. The form of the filter function determines the effects of the operator. A low-pass filter attenuates high frequencies, and retains low frequencies, so is used for smoothing and noise reduction. A high-pass filter yields edge enhancement or edge detection in the real domain (because edges contain many high frequencies), and therefore suppresses areas of constant grey level (which consists mainly of low frequencies). A band pass filter combines both of these functions. A suitable choice of filter enables the vessel pattern to be extracted. Further analysis can then be carried out by extracting shape descriptors that quantise the branching pattern of the vessels. These techniques can also be of use in detection of the optic disk.

Optic Disc Boundary Detection

The optic disk (OD) appears as a pale circular region in the human retina. It is the entry point of nerves and blood vessels into the retina. Its size and morphology are indicative of the general health of the retina, and distortions are indicative of disease leading to blindness. Many structures of the human retina are positioned in known relation to the optic disk.

The detection of the OD is not trivial, as the region appears bright, but crossed by dark blood vessels. This complicates identification of the boundary, which becomes fragmented. In addition, the size and the colour of the optic disk vary between individuals.

The method we have used for optic disk detection uses a high-pass filter to extract the edge of the OD, followed by a grayscale morphological closing operation with a flat disk-shaped structuring element. This is followed by applying Canny edge detection (Jelinek, Depardieu, Lucas, Cornforth, Huang, & Cree, 2005). We achieved an accuracy of 70% (14 correct out of 20 images), which can certainly be improved by adjusting the edge detection parameters, but compares favorably to previous work in this area. However it is to be noted that the result is excellent with respect to processing time and simplicity of the method (Figure 1).

Once the optic disk has been detected, the size and shape can be estimated, or it can be used as the starting point for vessel tracking, in order to determine the diameters of arterioles to venules, or it can be removed from the image by subtraction to simplify processing of the rest of the image.

ECG Assessment

In contrast to analysis of the retina, ECG assessment relies on processing of time series signals, of which an example is shown in Figure 2. Several automated classification systems for ECG assessment are in use, with the Minnesota code being the early de facto standard. This was followed by the NOVACODE (Rautaharju, Park, Chaitman, Rautaharju, & Zhang, 1998). For epidemiological studies, however, several individual ECG categories can be pooled into minor and major groupings to enhance statistical analysis, and are still in use (Klein, Klein, McBride, & Moss, 2005).

Figure 1. Retinal images showing application of the canny edge detector; left: good detection; right: detection with two errors

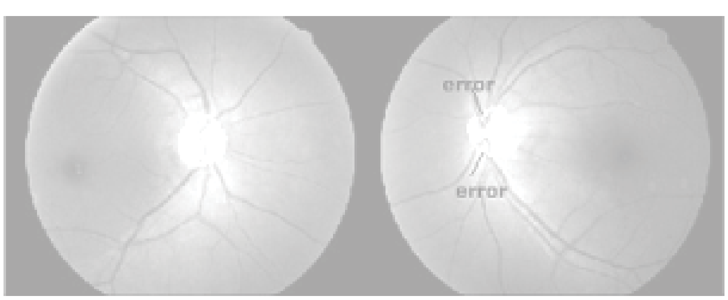


Figure 2. A typical three-lead ECG showing R peaks

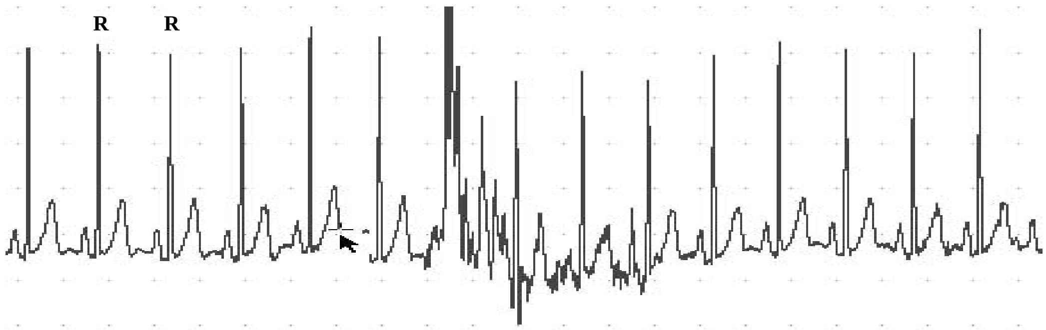
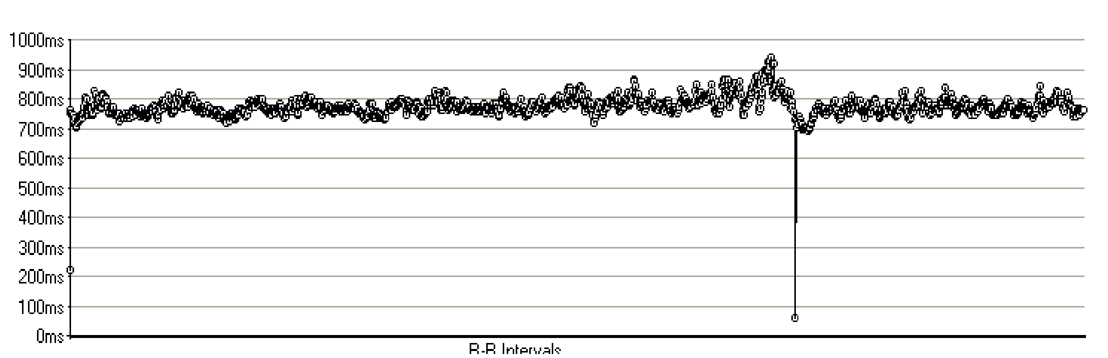


Figure 3. The R-R intervals are quantised and plotted against time to indicate HRV



The Charles Sturt Diabetes Complications Group (CSDCG) has investigated the possibility of primary health care providers assessing heart function using an automated ECG assessment tool that differs from the traditional method, by evaluating HRV using a three-lead recording. This recording and interpretation tool adds two important features to heart assessment. First, it allows primary health care providers to record ECGs, as placing three-lead ECGs is simpler compared to placing a 12-lead ECG, and less invasive in community health settings. Second, it is able to identify asymptomatic changes in heart function associated with diabetes. HRV is based on observations that many cardiac anomalies including CAN, change the HRV before any overt ECG anomalies are detectable. Cardiologists with current methods cannot pick up preclinical cardiac anomalies, but newer methods such as HRV have the potential to do this. In addition, relatively short three-lead ECG recordings can be used in evidence-based health care and accountability for primary health care practice, showing a similar efficacy in identifying at risk individuals as 12-lead ECG recordings.

Heart Rate Variability

To assess CAN, the variation in the interval between successive beats can be analyzed in terms of HRV. An ECG obtained using the three-lead method is shown in Figure 2. The large peaks are referred to as the R portion of the ECG, and the time between subsequent R peaks gives the period of the ECG. From the raw ECG recording the R-R intervals are obtained as shown in Figure 3, and analysed.

HRV analysis has been shown to be effective in diverse clinical situations including risk assessment of future myocardial infarction, changes associated with hypertension, Parkinson's disease, eating disorders, and diabetes (Gerritsen, Dekker, TenVoorde, Kostense, Heine, Bouter, Heethaar, & Stehouver, 2001; Kallio, Haapaniemi, Turkka, Suominen, Tolonen, Sotaniemi, Heikkila, & Myllyla, 2000; Mont, Castro, Herreros, Pare, Azqueta, Magrina, Puig, Toro, & Brugada, 2003).

A study on automated classification of diabetes, based on HRV, used data from the CSDCG study, which consisted of 46 adult people, 22 with known diabetes, and 24 controls, which have not been diagnosed with diabetes (Cornforth, Jelinek, Teich, & Lowen, 2004). The feasibility of classification was investigated using a variety of classifiers. It appears that diabetes could be detected from an analysis of HRV, but further study is obviously required. Diagnosis of diabetes by this method is not 100% accurate, and depends upon careful selection both of the classifier algorithm used, and the feature subset. As each classifier has different properties, there may be some further benefit by combining the results of classifiers.

Nonlinear Analysis of HRV

HRV recordings contain an irregular (aperiodic) component in addition to the low frequency and high frequency components. Detrended fluctuation analysis (DFA) is a form of nonlinear analysis, which best allows for the detection of long-range correlations. DFA is able to distinguish between people with congestive heart failure (CHF) and controls, and is a prognostic indicator for CHF that may complement traditional HRV measures in assessing cardiac risk in patients with CHF (Ho, Moody, Peng, Mietus, Larson, Levy, & Goldberger, 1997).

The majority of studies conducted using DFA involve recordings of two hours or greater. However, more recent studies showed the utility of shorter ECG recordings (Flynn, Jelinek, & Smith, 2005). Although CAN may be present in people with no diabetes, the study group was a sample of people with diabetes attending the CSDCG study. All participants were screened for CAN using the HRV 30:15 ratio. DFA analysis determines the long range correlations between successive R-R intervals. Table 1 shows that controls can be distinguished from people with diabetes and from those with CAN. Although there was no significant difference between those with and without CAN, there was a trend for the mean DFA value to be lowest for people with diabetes and CAN, followed by people with diabetes but without CAN, and highest for the control group.

Table 1. Results of a t-test on people with diabetes, diabetic autonomic neuropathy (DAN) and CAN shows the power of DFA to distinguish the early signs of diabetes

Comparison	**p**	**significant?**
Control / diabetes	0.03	yes
CAN^{+}/ CAN^{-}	0.41	no
Control / DAN^{+}	0.02	yes

The benefit of nonlinear analysis is that it provides a numeric scale that allows identification of the extent of cardiac dysfunction using short ECG recordings, making the results easy to interpret for nonspecialists, such as general practitioners. Also, as this numeric scale is an objective outcome measure, it can be used to determine the efficacy of treatment and ensure best practice.

Considering cardiovascular complications in diabetes account for 65% of all diabetic deaths, automated screening using nonlinear analyses has the potential to reduce cardiac morbidity through the early identification of pathology changes in people with diabetes. Nonlinear analysis also provides a means of following any progression in pathology associated with diabetes or otherwise that leads to cardiac dysfunction.

FUTURE TRENDS

It is already clear that automated procedures based on nonfluorescein photography of retinal images, and ECG based on the simple three-lead ECG, hold potential for the early detection of diabetes. What is not yet clear is how these techniques can be combined. We would expect that in the next few years, there will be attempts to perfect these analysis methods, and to apply data fusion, in order to combine the data from these sources. This will of course only be effective if applied to a routine screening of the general population. This depends on raising the awareness of diabetes among people in the community, especially among those in rural and regional areas, as this is where the greatest incidence of the disease occurs. The greatest obstacle to overcome in development of these new techniques will not be the technical aspects, that are already well on their way to being perfected, but rather in community acceptance.

CONCLUSION

Diabetes has been well recognized as a significant threat to health in many regions. Although treatment options are available, it remains a chronic disease contributing significantly to morbidity and mortality. Detection of the disease during its early stages holds promise for better treatment, higher quality of life, and a reduced burden on health provider services. However, this remains difficult.

Recent advances in computational methods provide a new range of tools that can be applied to early detection of diabetes. Among these are the automated analysis of retinal images and HRV, and the concept of a mobile "one-stop shop" for diabetes screening in rural communities. It is likely that these concepts will be further developed and integrated, in order to provide a rapid and reliable diagnostic tool that will be used as part of screening.

REFERENCES

Colagiuri, S., Colagiuri, R., & Ward, J. (1998). *National diabetes strategy and implementation plan.* Canberra Diabetes Australia: Paragon Printers.

Cornforth, D.J., Jelinek, H.F., Teich, M.C., & Lowen, S.B. (2004). Wrapper subset evaluation facilitates the automated detection of diabetes from heart rate variability measures. In M. Mohammadian (Ed.), *Proceedings of the International Conference on Computational Intelligence for Modelling Control and Automation (CIMCA'2004)* (pp. 446–455). University of Canberra, Australia.

Cree, M.J., Olson, J.A., McHardy, K.C., Forrester, J.V., & Sharp, P.F. (1996). Automated microaneurysms detection. In *Proceedings of the IEEE International Conference on Image Processing* (pp. 699–702). Lausanne, Switzerland: IEEE Press.

Dua, S., Kandiraju, N., & Thompson, H.W. (2005). Design and implementation of a unique blood-vessel detection algorithm towards early diagnosis of diabetic retinopathy. In *Proceedings of the International Conference on Information Technology: Coding and Computing* (Vol. 1, No. 1, pp. 26–31).

Flynn, A.C., Jelinek, H.F., & Smith, M.C. (2005). Heart rate variability analysis: A useful assessment tool for diabetes associated cardiac dysfunction in rural and remote areas. *Australian Journal of Rural Health, 13*, 77–82.

Gerritsen, J., Dekker, J.M., TenVoorde, B.J., Kostense, P.J., Heine, R.J., Bouter, L.M., Heethaar, R.M., & Stehouver, C.D.A. (2001). Impaired autonomic function is associated with increased mortality, especially in subjects with diabetes, hypertension or a history of cardiovascular disease. *Diabetes Care, 24*(10), 1793–1798.

Ho, K.K.L., Moody, G.B., Peng, C.K., Mietus, J.E., Larson, M.G., Levy, D., & Goldberger, A.L. (1997). Predicting survival in heart failure case and control subjects by use of fully automated methods for deriving nonlinear and conventional indices of heart rate dynamics. *Circulation, 96*(3), 842–848.

Jelinek, H.F., Cree, M.J., Worsely, D., Luckie, A.P., & Nixon, P. (2006). An automated microaneurysm detector as a tool for identification of diabetic retinopathy in a rural optometry practice. *Clinical and Experimental Optometry, 89*(5), 299–305.

Jelinek, H.F., Depardieu, C., Lucas, C., Cornforth, D., Huang, W., & Cree, M.J. (2005). Towards vessel characterisation in the vicinity of the optic disk in digital retinal images. In B. McCane (Ed.), *Proceedings of the Image and Vision Computing Conference*. University of Otago: New Zealand.

Kallio, M., Haapaniemi, T., Turkka, J., Suominen, K., Tolonen, U., Sotaniemi, K., Heikkila, V.P., & Myllyla, V. (2000). Heart rate variability in patients with untreated Parkinson's disease. *European Journal of Neurology, 7*, 667–672.

Klein, B.E., Klein, R., McBride, P.E., & Moss, S.E. (2005). Electrocardiographic abnormalities in individuals with long-duration type 1 diabetes. *Diabetes Care, 28*(1), 145–148.

Lee, S.C., Lee, E.T., Kingsley, R.M., Wang, Y., Russell, D., Klein, R., & Warn, A. (2001). Comparison of diagnosis of early retinal lesions of diabetic retinopathy between a computer system and human experts. *Archives of Ophthalmology, 119*(4), 509–515.

Mathers, C., & Penm, R. (1999). Health system costs of cardiovascular diseases and diabetes in Australia 1993–1994. In *Health and Welfare Expenditures* (pp 1–68). Canberra: Australian Institute of Health and Welfare.

Mont, L., Castro, J., Herreros, B., Pare, C., Azqueta, M., Magrina, et al. (2003). Reversibility of cardiac abnormalities in adolescents with anorexia nervosa after weight recovery. *Journal of the American Academy of Child & Adolescent Psychiatry, 42*(7), 808-813.

Rautaharju, P., Park, L.P., Chaitman, P.R., Rautaharju, F., & Zhang, Z.M. (1998). The novacode criteria for classification of ECG abnormalities and progression and regression. *Journal of Electrocardiology, 31*, 157–187.

Singalavanija, A., Supokavej, J., Bamroongsuk, P., Sinthanayothin, C., Phoojaruenchanachai, S., & Kongbunkiat, V. (2006). Feasibility study on computer-aided screening for diabetic retinopathy. *Japanese Journal of Ophthalmology, 50*(4), 361–6.

Soares, J.V.B., Leandro, J.J.G., Cesar, R.M. Jr., Jelinek, H.F., & Cree, M.J. (2006). Vessel segmentation using the 2-D Gabor wavelet and supervised classification. *IEEE Transactions in Medical Imaging, 25*(9), 1214–1222.

Wong, T.Y., Shankar, A., Klein, R., Klein, B.E., & Hubbard, L.D. (2005). Retinal arteriolar narrowing, hypertension and subsequent risk of diabetes mellitus. *Archives of Internal Medicine, 165*(9), 1060–1065.

KEY TERMS

Diabetic Retinopathy (DR): Damage to the eye caused by diabetes.

Fluorescein: Using fluorescent dye.

Microaneurysm: A bulge in a small blood vessel.

Computer Analysis of Coronary Doppler Flow Velocity

Valentina Magagnin
Politecnico di Milano, Italy
Istituto Ortopedico IRCCS Galeazzi, Italy

Maurizio Turiel
Università degli Studi di Milano, Italy
Instituto Ortopedico IRCCS Galeazzi, Italy

Sergio Cerutti
Politecnico di Milano, Italy

Luigi Delfino
Instituto Ortopedico IRCCS Galeazzi, Italy

Enrico Caiani
Politecnico di Milano, Italy
Istituto Ortopedico IRCCS Galeazzi, Italy

INTRODUCTION: THE CORONARY FLOW RESERVE

The coronary flow reserve (CFR) represents an important functional parameter to assess epicardial coronary stenosis and to evaluate the integrity of coronary microcirculation (Kern, 2000; Sadamatsu, Tashiro, Maehira, & Yamamoto, 2000). CFR can be measured, during adenosine or dipyridamole infusion, as the ratio of maximal (pharmacologically stimulated) to baseline (resting) diastolic coronary blood flow peak. Even in absence of stenosis in epicardial coronary artery, the CFR may be decreased when coronary microvascular circulation is compromised by arterial hypertension with or without left ventricular hypertrophy, diabetes mellitus, hypercholesterolemia, syndrome X, hypertrophic cardiomyopathy, and connective tissue diseases (Dimitrow, 2003; Strauer, Motz, Vogt, & Schwartzkopff, 1997). Several methods have been established for measuring CFR: invasive (intracoronary Doppler flow wire) (Caiati, Montaldo, Zedda, Bina, & Iliceto, 1999b; Lethen, Tries, Brechtken, Kersting, & Lambertz, 2003a; Lethen, Tries, Kersting, & Lambertz, 2003b), semi-invasive and scarcely feasible (transesophageal Doppler echocardiography) (Hirabayashi, Morita, Mizushige, Yamada, Ohmori, & Tanimoto, 1991; Iliceto, Marangelli, Memmola, & Rizzon, 1991; Lethen, Tries, Michel, & Lambertz, 2002; Redberg, Sobol, Chou, Malloy, Kumar, & Botvinick, 1995), or extremely expensive and scarcely available methods (PET, SPECT, MRI) (Caiati, Cioglia, Montaldo, Zedda, Rubini, & Pirisi, 1999a; Daimon, Watanabe, Yamagishi, Muro, Akioka, & Hirata, 2001; Koskenvuo, Saraste, Niemi, Knuuti, Sakuma, & Toikka, 2003; Laubenbacher, Rothley, Sitomer, Beanlands, Sawada, & Sutor, 1993; Picano, Parodi, Lattanzi, Sambuceti, Andrade, & Marzullo, 1994; Saraste, Koskenvuo, Knuuti, Toikka, Laine, & Niemi, 2001; Williams, Mullani, Jansen, & Anderson, 1994), thus their clinical use is limited (Dimitrow, 2003). In addition, PET and intracoronary Doppler flow wire involve radiation exposure, with inherent risk, environmental impact, and biohazard connected with use of ionizing testing (Picano, 2003a).

In the last decade, the development of new ultrasound equipments and probes has made possible the noninvasive evaluation of coronary blood velocity by Doppler echocardiography, using a transthoracic approach. In this way, the peak diastolic coronary flow velocity reserve (CFVR) can be estimated as the ratio of the maximal (pharmacologically stimulated) to baseline (resting) diastolic coronary blood flow velocity peak measured from the Doppler tracings. Several studies have shown that peak diastolic CFVR, computed in the distal portion of the left anterior descending (LAD) coronary artery, correlates with CFR obtained by more invasive techniques. This provided a reliable and noninvasive tool for the diagnosis of LAD coronary artery disease (Caiati et al., 1999b; Caiati, Montaldo, Zedda, Montisci, Ruscazio, & Lai, 1999c; Hozumi, Yoshida, Akasaka, Asami, Ogata, & Takagi, 1998; Koskenvuo et al., 2003; Saraste et al., 2001).

Moreover, this parameter has been shown to be able to detect coronary stenosis (> 90%) in the LAD coronary artery (Voci, Pizzuto, Mariano, Puddu, Chiavari, & Romeo, 2002) with high sensitivity and specificity (90% and 93%, respectively). In particular, a value of peak diastolic CFVR<1.9 has been associated to severe coronary stenosis (Matsumura, Hozumi, Watanabe, Fujimoto, Sugioka, & Takemoto, 2003), while a peak diastolic CFVR≥2.5 is representative of normal coronary epicardial stream (Picano, 2003b).

BACKGROUND: TRANSTHORACIC DOPPLER ECHOCARDIOGRAPHY

Transthoracic Doppler echocardiography is usually performed using an ultrasound unit equipped with a broad-band, high-frequency (3.5 to 7 MHz) transducer, with the patients in the left lateral decubitus position. The conventional imaging approach consists first in obtaining a short axis of the left ventricular apex and anterior groove to search for middistal LAD coronary flow using the Color-Doppler flow modality. When a diastolic circular color-coded blood flow is visualized in the anterior groove area, the transducer is rotated clockwise to obtain the best long axis of color flow, as visualized in Figure 1. Alternatively, a modified foreshortened two-chamber view could be obtained by sliding the transducer superiorly and medially from an apical two-chamber position.

Then, a careful search for color-coded blood flow is made over the epicardial part of the anterior wall with simultaneous attempts to optimize the visualization of the anterior groove area by very slight counterclockwise rotation and medial angling of the probe.

To minimize artifacts due to systolic cardiac contraction, the sample volume is placed in correspondence to the diastolic position of the explored vessel cross-section, and maintained fixed for all the acquisition period. Pulsed-wave Doppler spectral tracings of LAD flow velocity is then recorded by fast Fourier transformation analysis (Caiati et al., 1999a).

Particular care has to be taken in maintaining the pulsed-wave Doppler aligned parallel to the blood flow. The achievement of this task can be confirmed by the fact that the spectral Doppler flow shows the characteristic biphasic flow pattern, with a larger diastolic component, and a smaller systolic one.

Once the image is optimized, multiple still-frame image acquisition is performed to obtain representative images of consecutive beats at baseline, digitally stored for off-line analysis. Then, the maximal stress condition is pharmacologically induced by injecting adenosine or dipyridamole. When dipyridamole is utilized (0.56 mg/kg i.v. in four minutes, and successively 0.28 mg/kg i.v. in two minutes, to induce vasodilation of the microcirculation), image acquisition is performed after two to four additional minutes from the drug subministration, depending on the patient response, with the same ultrasound settings of the baseline acquisition (Lim, Shim, Rhee, Kim, Hwang, & Kim,

Figure 1. Example of a diastolic frame acquired with the color–doppler flow modality activated

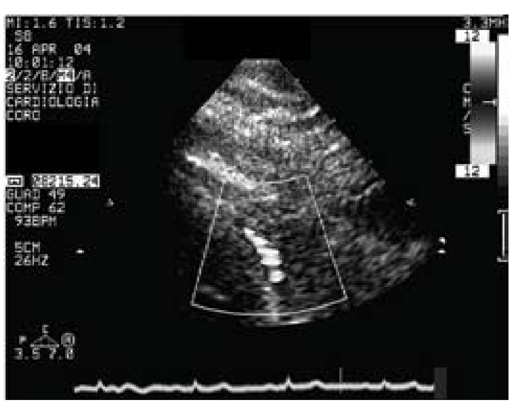
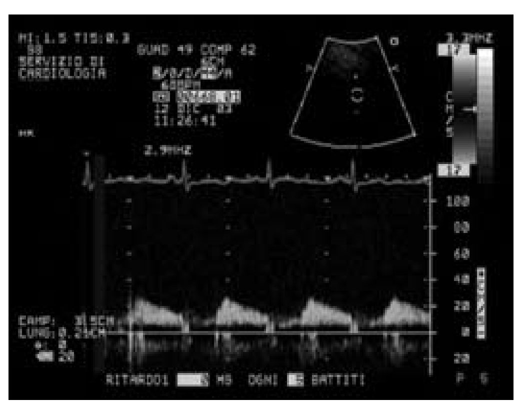


Figure 2. Example of the Pulsed-wave Doppler spettral tracings of flow velocity in the LAD coronary artery. The characteristic biphasic flow pattern can be observed.



2000), and, at the end of the protocol, 125 to 250 mg of aminophylline are subministered to counteract the effect of dipyridamole. In addition, ECG is continuously monitored on the ultrasound screen (one lead) and by conventional 12-leads ECG.

Background: Transthoracic Doppler Echocardiography Limitations

Despite the advantages introduced by the transthoracic Doppler echocardiography, in clinical practice the off-line analysis of the Doppler flow velocity is based on manual identification, often performed on a single beat, of few fiducial points (systolic and diastolic peak velocities). From these points, both peak diastolic CFVR and the diastolic-to-systolic velocity ratio (DSVR) are computed (Picano, Sicari, & Varga, 1999). This time-consuming manual analysis introduces subjectivity in the measurements of the extracted clinical parameters, as well as limits the information which could be exploited from the coronary Doppler velocity signal with a more automated analysis.

As transthoracic coronary Doppler echocardiography is rapidly gaining appreciation as a popular tool for the noninvasive evaluation of peak diastolic CFVR, we hypothesized that the computerized analysis of the coronary Doppler velocity signal would reduce the subjectivity in the results and would allow the calculation of additional parameters, which could be potentially useful in the diagnostic process. Accordingly, we developed a technique for nearly-automated detection of Doppler coronary flow velocity profile, and validated this technique in comparison with the conventional measurements obtained by manual tracings.

COMPUTERIZED ANALYSIS OF CORONARY FLOW VELOCITY PROFILE

One hundred patients (67 males, mean age 62±9), referred to transthoracic echocardiographic examination at the Cardiology Unit of the Ortophedic Galeazzi Hospital in Milan, were retrospectively analyzed. All patients were studied by a Sonos 5500 (Philips) ultrasound machine equipped with a S8 high-frequency probe, according to the standard protocol described in the previous paragraph, using dipyridamole.

Image Preprocessing

To automatically quantify parameters relevant to the Doppler flow velocity profile, a specific custom software was developed using Matlab (The MathWorks, Inc.; Natick, Massachusetts).

The digitally acquired images were automatically cropped to extract the region of interest (ROI) containing the Doppler tracings, with the relevant time and velocity scales, and the ECG signal. A horizontal Sobel filter was then applied to automatically detect the zero velocity line, thus cropping the ROI to contain only positive velocity Doppler tracings. In fact, the negative velocity components of the Doppler tracings are not determinant in the calculation of the peak diastolic CFVR.

To properly scale the measurements which will be computed, otherwise expressed in pixels, the vertical velocity axis and its horizontal ticks were detected by Sobel filtering. The distance between two consecutive ticks was automatically computed and associated with the scale factor, manually inserted by the user.

To properly obtain the temporal information relevant to the duration of each consecutive beat, imaged in the same still-frame, the ECG signal, visualized in green at the top part of the ROI, was extracted exploiting the colour information. To prevent unambiguous detection of the R-wave peak for every cardiac cycle, a skeletonization operator was applied to thin the ECG trace at one pixel thickness. Finally, the fiducial points corresponding to the R-wave peaks were detected by a conventional threshold-derivative algorithm.

Detection of the Doppler Velocity Profile

To automatically detect the Doppler velocity profile from the Doppler tracings, a local adaptive thresholding algorithm was developed and applied. First, the ROI was divided in three partially overlapped equally sized rectangular regions. For each ROI, its inverse normalized cumulative videointensity histogram was computed as $\%\sum_{i=I}^{256} \#\, pixel(i)$.

The videointensity L, at which 25% of the pixels had an intensity greater than L, was assumed to be the threshold for that region (see Figure 3).

In the overlapped areas, the mean of the thresholds relevant to the two adjacent regions was utilized. The value of 25% was preliminary determined as the optimal choice as follows: first, the images acquired from a randomly selected subset of 20 patients, both at baseline and during stress, were analyzed with the proposed local adaptive threshold algorithm using different presettings for the automated threshold (from 10% to 40%, with a 5% step). Then, each automatically detected Doppler velocity profile contour was shown in random order, overimposed to the original image, to two expert cardiologists for visual evaluation. The cardiologists, blinded to the threshold value, had to reach a consensus on which of the presented images better visualized the true Doppler velocity profile contour. The images corresponding to the contour extraction using the 25% threshold value were selected in 90% of the cases.

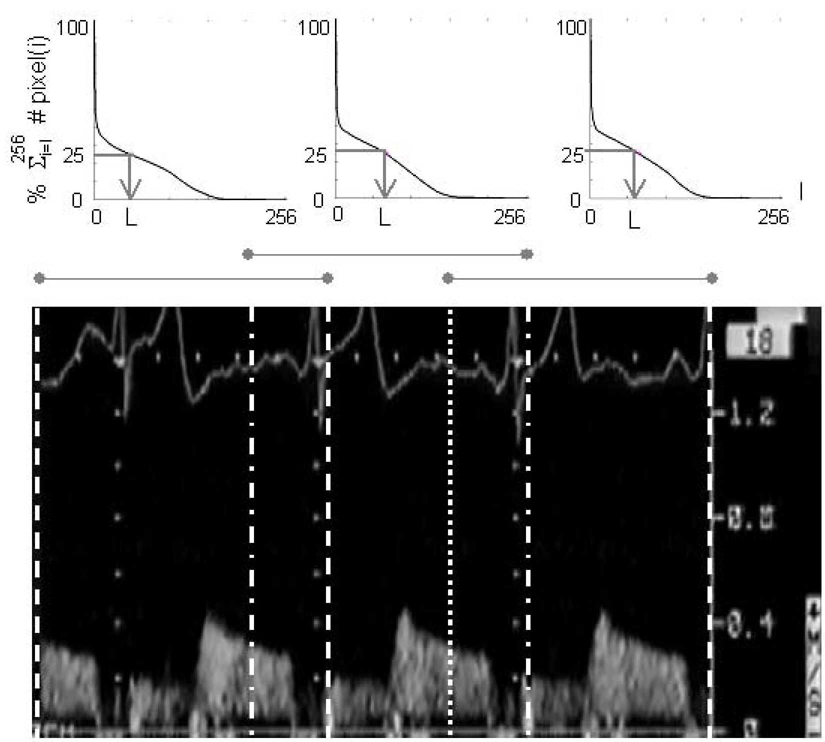
Each pixel column was scanned from the bottom to the top, to search for the first pixel with intensity lower than the corresponding threshold, thus obtaining a binary image representing the detected Doppler velocity profile. Finally, the 1-D velocity signal was extracted and filtered (median filter, order 15) to remove outliers and artefacts, and then overimposed to the ROI to allow visual verification of the reliability of the detected profile, and eventual manual correction of the thresholds. Subsequently, the detected Doppler profile was interpolated (cubic spline) to obtain a 1000 samples/sec oversampling; then, a low-pass FIR filter (150 coefficients, cut-off frequency 30 Hz) was applied, in order to reduce high frequency noise, and the first derivative was computed.

Parameters Extraction

In order to automatically obtain conventional clinical parameters from the detected Doppler velocity profile, some fiducial points needed to be identified in consecutive beats. From the detection of the R-wave peak positions, the beginning (R(i)) and ending (R(i+1)) points of the Doppler profile relevant to each consecutive cardiac cycle were automatically identified. Only the cardiac cycles that were completely represented in the still-frame image were considered in the further analysis.

Basing on the a-priori knowledge of the expected velocity profile morphology (biphasic pattern, with a systolic and a diastolic velocity components), for each

Figure 3. Bottom: doppler flow velocity profile image divided in three partially overlapped rectangular regions (black lines); top: the inverse normalized cumulative videointensity histogram, and the videointensity L at which the 25% of the pixels had an intensity greater than L, is shown for each region



C

cardiac cycle the following fiducial points were identified (see Figure 4): (1) peak diastolic velocity (PDV), as the absolute maximum in a temporal window ranging from ¼ to ¾ of the RR duration; (2) peak diastolic acceleration (PDA), as the first derivative absolute maximum in a temporal window ending at PDV, with a duration equal to 20% of the total beat duration; (3) the beginning of the diastolic phase (BD), as the absolute minimum of velocity in a temporal window ending at PDA, with a duration equal to 15% of the total beat duration; (4) peak systolic velocity (PSV), as the absolute maximum in the window ranging from R(i) to BD; (5) peak diastolic deceleration (PDD), as the derivative absolute minimum in a temporal window starting from PDV to R(i+1).

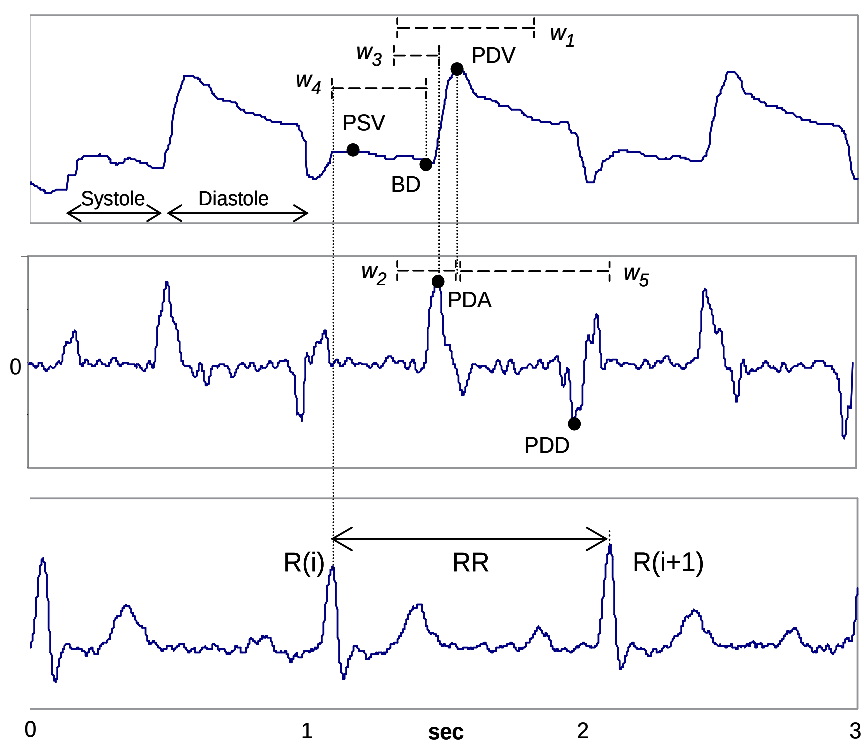
From these fiducial points, the following clinical measurements can be obtained: systolic and diastolic peak velocities (PSV and PDV), DSVR, peak diastolic CFVR. All the measurements obtained from three consecutive cardiac cycles were automatically averaged to reduce measure variability, and to take into account possible respiratory variations.

Validation Protocol

For each subject, still-frame images, including three to four cardiac cycles relevant to both baseline and stress, were independently analyzed by two observers (A1 and A2), using the developed custom software. To obtain a reference for comparison, the same images were manually evaluated by two expert cardiologists (M1 and M2), blinded to each other and to the results of the automated analysis, to extract the conventional clinical indices of PDV, PSV, peak diastolic CFVR and DSVR. For each index, the mean value was considered as the "independent standard" value.

The mean of the computerized measurements, obtained by A1 and A2, was compared to the manual "independent standard" by linear regression and Bland-Altman analyses. For each linear regression, the correlation coefficient was computed ($p<0.05$). For Bland-Altman analysis, the confidence interval was defined as ± 2SD, and the significance of the bias was tested by a paired t-test vs. null values ($p<0.05$). For both manual and automated measurements, the

Figure 4. Schematization of the automated detection of fiducial points from the doppler velocity profile (top), its first derivative (middle) and the ECG (bottom)



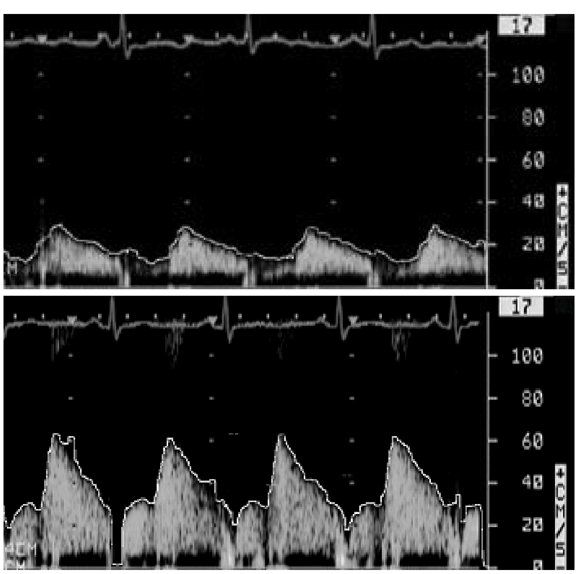
inter-observer variability was computed as the SD of the two respective measurements, and expressed as percentage of their mean.

Results

A set of 200 Doppler velocity images (100 at baseline and 100 at stress) from 100 patients was analyzed. In these images, the horizontal resolution was fixed and equal to 128 pixels/sec, while the vertical resolution varied according to the velocity scale settings, in a range between 150 and 300 pixels/m/sec.

The proposed method was able to correctly detect the Doppler velocity profile in 180/200 of the analyzed frames without the need of manual interaction, using the preset algorithm parameters. In the remaining 20/200, a manual adjustment of the threshold parameter was necessary to correctly identify the profile. The required time for analysing a single frame containing three consecutive beats ranged between 10 sec. (with automated threshold setting) and 40 sec (with manual correction of threshold setting), using a Pentium M 745 (1.8 GHz CPU) personal computer, including data retrieval. Conversely, conventional manual analysis of a single frame required about 5 min., including data retrieval and mean computations.

Figure 5. Doppler flow velocity images with good image quality, visualized with overimposed the corresponding automatically detected velocity profile contour; top: baseline; bottom: stress



In Figure 5, examples of the detected Doppler velocity profile obtained using the preset 25% threshold value are shown.

The results of the comparisons with the "independent standard" values showed excellent correlation in PSV (y=1.01x-0.17, r=0.986, p<0.0001), PDV (y=1.02x-0.86, r=0.998, p<0.0001), and peak diastolic CFVR (y=0.97x+0.07, r=0.987, p<0.0001). Also, a good correlation was found in DSVR (y=x+0.02, r=0.809, p<0.0001).

Bland-Altman analysis resulted in not significant biases (i.e., mean of the signed differences) and small confidence intervals in each of the examined parameters (bias±2SD, PSV: -0.2±5.0 cm/s; PDV: -0.4±3.2 cm/s; CFVR: 0.01±0.25; DSVR: 0.0±0.56).

Inter-observer variability between A1 and A2, resulting from the computed indices using the proposed computerized method, was low (baseline: PSV: 6.84±8.6 %; PDV: 1.65±2.4 %; DSVR: 6.64±9.4 %; stress: PSV: 3.71±5.2 %; PDV: 1.90±2.4 %; DSVR: 4.69±8.3 %; peak diastolic CFVR: 2.65±3.3 %) and comparable with that resulting from M1 and M2 conventional manual tracings (baseline: PSV: 3.77±3.8 %; PDV: 2.53±2.3 %; DSVR: 4.23±4.7 %; and stress: PSV: 5.20±7.0 %; PDV: 1.98±2.0 %; DSVR: 5.78±6.4 %; peak diastolic CFVR: 3.16±2.6 %).

DISCUSSION

The possibility to noninvasively obtain information about coronary blood velocity by transthoracic Doppler echocardiography opens new scenarios in the diagnosis and treatment of coronary artery disease. However, the need for manual analysis of the Doppler tracings introduces a limitation to the clinical applicability of this methodology. To overcome this limitation, computerized methods for analysis of the Doppler coronary flow velocity profile are needed. Accordingly, we developed an adaptive thresholding algorithm, capable to take into account artifacts (high frequency or speckle noise) potentially existing in the Doppler coronary flow velocity signal, and extract clinically useful parameters.

The choice to divide the image in three partially overlapping regions, and to compute for each region, based on its histogram, the videointensity value to be assumed as the local threshold, allowed the velocity profile detection to be reliable, even in presence of low quality images. The choice of the preset optimal dimension of the median filter was defined in order to obtain a good compromise between accuracy of the detection and suppression of local artefacts. When compared with manual tracing, the proposed computerized approach reduced the time necessary for the analysis, thus preserving the reliability of the results and their reproducibility. In fact, both linear regression and Bland-Altman analyses confirmed the high correlation, absence of bias, and narrow limits of agreement, pointing out the excellent correlation between the results obtained with both the techniques and the potential clinical applicability of the proposed method. Also, the interobserver variability measured in the clinical parameters by the proposed technique was low and similar to the results obtained by manual analysis. The largest variability was observed in DSVR: this parameter, representing the ratio between the peak diastolic and systolic velocities, probably reflects and amplifies the higher variability observed in the peak systolic velocity.

The availability of a reliable digitized Doppler flow velocity profile could allow in the future to extended the routine clinical analysis to the computation of additional amplitude and temporal parameters (i.e., mean velocities, phase durations, diastolic slope, peak derivatives) other than the conventional indices.

CONCLUSION

In conclusion, the measurements obtained with the proposed computerized approach were found accurate and reproducible compared with manual tracing, with a reduction of the time needed for the analysis.

The proposed method allows the computation of additional indices, which may be useful in the diagnostic phase. Moreover, it could potentially be applied to evaluate the coronary flow velocity measured by intracoronary devices (Doppler flow wire), or the Doppler flow velocity recorded in other arteries (i.e., brachial, femoral) or districts (aortic or mitral inflow). However, the preliminary phase of optimal threshold value assessment should be repeated for each specific application, to confirm if the 25% value could still be considered as the optimal one.

REFERENCES

Caiati, C., Cioglia, G., Montaldo, C., Zedda, N., Rubini, G., Pirisi, R. et al. (1999a). Correlation of 99mTc-sestamibi SPECT myocardial perfusion imaging with absolute coronary flow reserve by a new noninvasive doppler method in patients with stenoses of the left anterior descending coronary artery. A preliminary report. *Cardiologia, 44*, 809–816.

Caiati, C., Montaldo, C., Zedda, N., Bina, A., & Iliceto, S. (1999b). New noninvasive method for coronary flow reserve assessment: Contrast-enhanced transthoracic second harmonic echo doppler. *Circulation, 99*, 771–778.

Caiati, C., Montaldo, C., Zedda, N., Montisci, R., Ruscazio, M., Lai, G. et al. (1999c). Validation of a new noninvasive method (contrast-enhanced transthoracic second harmonic echo doppler) for the evaluation of coronary flow reserve: Comparison with intracoronary doppler flow wire. *Journal of the American College of Cardiology, 34*, 1193–1200.

Daimon, M., Watanabe, H., Yamagishi, H., Muro, T., Akioka, K., Hirata, K. et al. (2001). Physiologic assessment of coronary artery stenosis by coronary flow reserve measurements with transthoracic doppler echocardiography: Comparison with exercise Thallium-201 single positron emission computed tomography. *Journal of the American College of Cardiology, 37*, 1310–1315.

Dimitrow, P.P. (2003). Transthoracic doppler echocardiography - noninvasive diagnostic window for coronary flow reserve assessment. *Cardiovascular Ultrasound, 1*, 4.

Hirabayashi, K., Morita, H., Mizushige, K., Yamada, S., Ohmori, K., Tanimoto, T. et al. (1991). Imaging of coronary arteries and measurement of coronary flow using transesophageal doppler echocardiography: A comparison with transthoracic doppler echocardiography. *Journal of Cardiology, Supplement, 26*, 15–24.

Hozumi, T., Yoshida, K., Akasaka, T., Asami, Y., Ogata, Y., Takagi, T. et al. (1998). Noninvasive assessment of coronary flow velocity and coronary flow velocity reserve in the left anterior descending coronary artery by doppler echocardiography: Comparison with invasive technique. *Journal of the American College of Cardiology, 32*, 1251–1259.

Iliceto, S., Marangelli, V., Memmola, C., & Rizzon, P. (1991). Transesophageal doppler echocardiography evaluation of coronary blood flow velocity in baseline conditions and during dipyridamole-induced coronary vasodilation. *Circulation, 83,* 61–69.

Kern, M. J. (2000). Coronary physiology revisited: Practical insights from the cardiac catheterization laboratory. *Circulation, 101,* 1344–1351.

Koskenvuo, J. W., Saraste, M., Niemi, P., Knuuti, J., Sakuma, H., Toikka, J. O. et al. (2003). Correlation of transthoracic doppler echocardiography and magnetic resonance imaging in measuring left anterior descending artery flow velocity and time-course of dipyridamole-induced coronary flow increase. *Scandinavian Journal of Clinical and Laboratory Investigation, 63,* 65–72.

Laubenbacher, C., Rothley, J., Sitomer, J., Beanlands, R., Sawada, S., Sutor, R. et al. (1993). An automated analysis program for the evaluation of cardiac PET studies: Initial results in the detection and localization of coronary artery disease using Nitrogen-13-ammonia. *Journal of Nuclear Medicine, 34,* 968–978.

Lethen, H., Tries, H. P., Brechtken, J., Kersting, S., & Lambertz, H. (2003a). Comparison of transthoracic doppler echocardiography to intracoronary doppler guidewire measurements for assessment of coronary flow reserve in the left anterior descending artery for detection of restenosis after coronary angioplasty. *American Journal of Cardiology, 91,* 412–417.

Lethen, H., Tries, P., Kersting, S., & Lambertz, H. (2003b). Validation of noninvasive assessment of coronary flow velocity reserve in the right coronary artery. a comparison of transthoracic echocardiographic results with intracoronary doppler flow wire measurements. *European Heart Journal, 24,* 1567–1575.

Lethen, H., Tries, H. P., Michel, R., & Lambertz, H. (2002). Transesophageal echocardiography. Technique and diagnostic possibilities. *Internist (Berl), 43,* 972–984.

Lim, H. E., Shim, W. J., Rhee, H., Kim, S. M., Hwang, G. S., Kim, Y. H. et al. (2000). Assessment of coronary flow reserve with transthoracic doppler echocardiography: Comparison among adenosine, standard-dose dipyridamole, and high-dose dipyridamole. *Journal of the American Society of Echocardiography, 13,* 264–270.

Matsumura, Y., Hozumi, T., Watanabe, H., Fujimoto, K., Sugioka, K., Takemoto, Y. et al. (2003). Cut-off value of coronary flow velocity reserve by transthoracic doppler echocardiography for diagnosis of significant left anterior descending artery stenosis in patients with coronary risk factors. *American Journal of Cardiology, 92,* 1389–1393.

Picano, E. (2003a). Stress echocardiography for the diagnosis of coronary artery disease. *Indian Heart Journal, 55,* 223–227.

Picano, E. (2003b). Stress echocardiography: A historical perspective. *American Journal of Medicine, 114,* 126–130.

Picano, E., Parodi, O., Lattanzi, F., Sambuceti, G., Andrade, M. J., Marzullo, P. et al. (1994). Assessment of anatomic and physiological severity of single-vessel coronary artery lesions by dipyridamole echocardiography. Comparison with positron emission tomography and quantitative arteriography. *Circulation, 89,* 753–761.

Picano, E., Sicari, R., & Varga, A. (1999). Dipyridamole stress echocardiography. *Cardiology Clinics, 17,* 481–99.

Redberg, R. F., Sobol, Y., Chou, T. M., Malloy, M., Kumar, S., Botvinick, E. et al. (1995). Adenosine-induced coronary vasodilation during transesophageal doppler echocardiography. Rapid and safe measurement of coronary flow reserve ratio can predict significant left anterior descending coronary stenosis. *Circulation, 92,* 190–196.

Sadamatsu, K., Tashiro, H., Maehira, N., & Yamamoto, K. (2000). Coronary microvascular abnormality in the reversible systolic dysfunction observed after noncardiac disease. *Japanese Circulation Journal, 64,* 789–792.

Saraste, M., Koskenvuo, J., Knuuti, J., Toikka, J., Laine, H., Niemi, P. et al. (2001). Coronary flow reserve: Measurement with transthoracic doppler echocardiography is reproducible and comparable with positron emission tomography. *Clinical Physiology, 21,* 114–122.

Strauer, B. E., Motz, W., Vogt, M., & Schwartzkopff, B. (1997). Impaired coronary flow reserve in NIDDM: A possible role for diabetic cardiopathy in humans. *Diabetes,* 119S–124S.

Voci, P., Pizzuto, F., Mariano, E., Puddu, P. E., Chiavari, P. A., & Romeo, F. (2002). Measurement of coronary flow reserve in the anterior and posterior descending coronary arteries by transthoracic doppler ultrasound. *American Journal of Cardiology, 90*, 988–991.

Williams, B. R., Mullani, N. A., Jansen, D. E., & Anderson, B. A. (1994). A retrospective study of the diagnostic accuracy of a community hospital-based PET center for the detection of coronary artery disease using Rubidium-82. *Journal of Nuclear Medicine, 35*, 1586–1592.

KEY TERMS

Bland-Altman Analysis: Statistical method which allows to compare measurements obtained in the same subjects by two different techniques. The Bland-Altman graph plots the difference between the two measurements against their average, for each subject.

Coronary Flow Reserve: Functional clinical parameter, computed as the ratio of maximal (pharmacologically stimulated) to baseline (resting) diastolic coronary blood flow peak, used to assess epicardial coronary stenosis and to evaluate the integrity of coronary microcirculation.

Dipyridamole: Drug that inhibits platelet aggregation and causes vasodilation. It is used in cardiac stress testing as a coronary vasodilator to induce the maximal stress condition.

Image Processing: The process of applying algorithms to images in order to improve their quality reducing artefacts, to enhance characteristics, to detect object contours.

Left Anterior Descending Coronary Artery: One of the two branches (together with the **circumflex (Circ)** coronary artery) in which the **left main** coronary artery divides into.

Local Adaptive Thresholding Algorithm: Technique based on the computation of a locally dependent threshold in a determined region of interest, by which the object contour is then detected by thresholding.

Thresholding: The process of detecting object contours based on the computation of a binary image in which pixel values are set to one, in correspondence to a videointensity value of the corresponding pixel position in the analyzed image greater than the threshold, and to zero if the videointensity is lower than the threshold. The threshold can be constant for all image, or can be adaptive and spatially dependent.

Transthoracic Doppler Echocardiography: Noninvasive evaluation of coronary blood velocity by Doppler echocardiography, using a transthoracic approach.

Costs and New Technologies in Healthcare Delivery

Marie-Christine Thaize Challier
University of Grenoble, France

INTRODUCTION

Reducing medical costs is one of the major policy questions. A well understanding of this issue requires specific insight into some domains. The three main points which must be clarified relate to: (1) the nature of the costs (production costs and transaction costs), (2) the nature of *new healthcare technologies* (the biotechnology and biomedical engineering vs. the e-healthcare system, the face-to-face communications between the primary care physicians or other providers and the patients vs. the non face-to-face communications), and (3) the nature of *information* (which can be asymmetric, incomplete, or imperfect). This article studies what differences there are when considerations about informational issues and types of medicine are taken into account in the modern health economy characterized by the generation and the implementation of new healthcare technologies. To do this, this article first clarifies and presents some concepts in the framework of the health economy. Then, from these issues, it discusses the cost containment in the current case of the development and adoption of new technologies, and contrasts the opinions and perspectives.

BACKGROUND: COSTS AND INFORMATION ISSUES

An approach for a better understanding of healthcare delivery requires particular insights into the nature and effect of new technologies, information, and costs.

Technologies, Information, and Costs

A first interest concerns the consequences of new medical technological innovations on medical costs. *(i)* On the one hand, biotechnologies (defined by technological applications using biological systems or living organisms to make or change products or processes), biomedical engineering (i.e., the application of engineering techniques and principles), and nanotechnologies (a variety of techniques to produce materials and devices on the scale of the nanometer, their societal implications being a controversial subject) are expensive. Nowadays, this healthcare industry represents one of the international largest and fastest-growing industries. Yet, accurate data on the genuine costs of such new healthcare technologies are not available. *(ii)* On the other hand, technological improvements based on e-healthcare technologies (over the Internet, e-mail, text messages, video conferences, and other ways) seem to reduce costs through efficient use of technology. For financial reasons, e-healthcare systems are becoming strategic necessities. Indeed, various networking technology and information have introduced major transformation in healthcare delivery in particular through telemedicine, electronic patient records, technology-mediated solutions, network design issues for healthcare operations, among others. Moreover, by introducing non face-to-face communications between the practitioners and the patients, the latter is empowered through better access to medical information and care.

The second point which deserves some details is relative to the term information. In the economic sense, it is usual to refer to three major concepts: asymmetric information, incomplete information, and imperfect information. *(i) Informational asymmetries* occur when one party to a transaction knows more information than the other parties either about his/her own position or about the environment. Such asymmetries in information lead to distortions when parties define strategies or assess a situation. Asymmetrical information can induce opportunistic behaviors and can be responsible for problems of contractual performance: the adverse selection problem and the moral hazard problem. "*Adverse selection*" refers to situations in which, *before* the transaction occurs, the ignorant party lacks information about some aspect concerning the characteristics of the other party or the object of the transaction (Akerlof, 1970). "*Moral hazard*" refers to situations where, *after* the transaction occurs, the informed party engages in activities that are undesirable from the other party (Arrow, 1963; Holmström, 1979). *(ii) Informational incompleteness* means that there is a lack of complete information between two parties, the actions or characteristics of one party being not observable by the other one. In other words, one party has private information about something relevant to his/her decision making. *(iii) Imperfect information* refers to situations where a party does not know the previous actions of the other party (Rothschild & Stiglitz, 1976). For example, a physician ignores the patients' behaviors with regard to their own health.

A third clarification concerns the tendency to adopt organizational modes that best reduce healthcare costs, specifically production costs and transaction costs. If the former are generally well-defined, the latter deserve more specific attention. Economists define transaction costs as the costs of running the system: costs of coordinating and of motivating (Milgrom & Roberts, 1992; Williamson, 1979). Coordination costs include, among others, search costs, communication costs, decision costs, and enforcement costs while motivation costs refer to informational situations and encompass costs associated with imperfect, incomplete, and asymmetric information (Milgrom & Roberts, 1992). Such transaction costs can be reduced in the three types of medicines: the preventive medicine (act of protecting, promoting, maintaining health, and preventing disease), the diagnostic medicine (act or process of identifying the cause and nature of the illness through the evaluation of patient history and the examination of laboratory data), and the curative medicine (which covers the actions and treatments correcting a harmful or troublesome situation).

Rising Costs and Medical Practices Faced to Moral Hazard, Supplier-Induced Demand, and Adverse Selection

Besides technological change and aging, most researchers and practitioners focus on three main causes of rising spending on medical care without proportionate increases in the quality of care: *moral hazard* and *supplier-induced demand*.

1. A first informational problem and source of inefficiency is due to "moral hazard," specifically here to an excessive demand by insured patients. Indeed, in this specific case, moral hazard reflects the state of mind and the change in behavior that can occur when a person becomes insured. Through their insurance, patients pay less out of their own pocket when they consume medical services, and consequently they can demand more of them. In other words, if insurance covers the full cost of care, policyholders may overuse medical services, leading to raise costs. Researchers distinguish *ex ante moral hazard*, which is relative to the absence or lack of preventive action by insured *and ex post moral hazard*, which refers to the insured behavior when the illness appears, such as medical nomadism—people seeking advice from various physicians for the same disease—or repetitive tests. Therefore, to prevent moral hazard, insurers—following the RAND Health Insurance Experiment—limit the incentives of patients to demand services. To do this, thinking that *price matters*, insurance companies share healthcare expenses by imposing a copayment, namely a fixed portion of the actual medical cost that an insured person must pay (Dranove, 2000, pp. 28–31). They use copayments in order to: (1) prevent unnecessary medical care, (2) rein in medical costs, and (3) lead to savings for insurance companies. However, this health-insurance policy, according to which insurers bear only an amount of the actual cost, can produce perverse effects. For example, copayments can cut appropriate and necessary office visits and medications. More broadly**,** they can reduce preventive and curative care (Keeler, Brook, Goldberg, Kamberg, & Newhouse, 1985; Lurie, Kamberg, Brook, Keeler, & Newhouse, 1989). Furthermore, insured people who make copayments can be healthier than those who do not, and therefore purchase fewer medical services; in that case, *price might not matter*.
2. A second informational problem and source of inefficiency is due to the phenomenon of "supplier-induced demand" (Evans, 1974; Pauly, 1994; Roemer, 1961). The demand inducement problem occurs when primary physicians use their superior knowledge and their influence to generate demand for personal gain, without health benefits. Due to this information gap between such physicians and patients, the former (who are both advisors to the patients and providers of services) can make

treatment recommendations and various clinical decisions that raise the costs, and are not in the best interests of their patients. It is worth emphasizing that this agency problem occurs because, not only physicians are more fully informed than patients, but also because they can have financial incentives to recommend costly treatments that are of little value to patients. Patient ignorance provides doctors with discretionary power to manipulate demand; clinical recommendations for medical treatment can be ambiguous, harming the patient's interest (Folland, Goodman, & Stano, 2001, chap. 10; Rice & Labelle, 1989). The supplier-induced demand is a controversial problem: while inducement theorists put forth that supply causes demand, other researchers develop different opinions or causalities. Reinhard (1989) considers as a matter of the utmost importance to know whether healthcare markets function according to the demand or supply side. Under fee-for-service payments, physicians can induce demand by overtreating their patients, who thus purchase more medical care; it results that such behaviors drive up medical costs, and increases unnecessary and inappropriate cares. Since the price paid by insured consumers when they apply for healthcare services can be set separately from the price paid to providers when services are supplied, the two alternative ways for controlling the spending are: *(i)* demand-side cost sharing (where patients must pay more in copayments), and *(ii)* supply-side cost sharing (in order to change the incentives of suppliers to provide certain services (for the development of such strategies, refer to Ellis & McGuire, 1993). Whatever the causalities, treatments seem more dependent on economic, rather than medical, criteria.

NEW MEDICAL TECHNOLOGIES FACED TO MEDICAL COSTS

Healthcare costs rise despite several techniques to limit them. New economic issues spring through the generation and implementation of new healthcare technologies and e-healthcare technologies; in particular, those developing solutions to cut healthcare costs and insurance costs.

Technological Change and Costs in the Healthcare System

Costs of Healthcare

Technological change may reduce costs when it increases the productivity of healthcare resources by providing less costly production methods for existing products (when less expensive inputs replace more costly inputs) or may increase costs when it promotes new and expensive products. Measuring the cost of a medical treatment when such "process innovations" and "product innovations" change the access to care, and the treatment is not easy. Scitovsky (1985) finds that technological change increases the cost of several diseases. Other researchers explain increasing costs of healthcare by highlighting the "medical arms race;" namely, the competition between hospitals for physicians and patients by supplying the latest medical technology and the largest staffing (Dranove, 2000; Robinson & Luft, 1985). Moreover, since physicians are both diagnosticians and clinicians, patients trust them to undertake the right decisions. Consequently, especially in the case of new treatments and technologies, patients are more vulnerable, having more difficulty to assess the value of the new care. This represents thus another source of inflation if physicians raise their incomes by unnecessary or inappropriate recommendations or other medical decisions. Such misbehaviors can be higher when physicians own diagnostic testing equipment (Dranove, 2000).

Costs of Health Insurance

Indemnity insurance is often considered as a source of the development of costly medical technology (Goddeeris, 1984a). Insurance coverage seems to increase costs through its weak direct effect on demand and its strong inducement of cost-increasing technological innovations. In this context, Newhouse (1988) and Peden and Freeland (1998) support the "induced-innovation hypothesis." Investigating the following question: does medical technological change make the average patient better off? Goddeeris (1984a, 1984b) finds that indemnity insurance increases the relative profitability of cost-increasaing technological innovations towards cost-cutting ones, insurance leading to "more costly mix of medical innovations than would otherwise occur" (pp. 530). He shows that the average patient pays the full cost of the new healthcare technologies since his/her insurance premiums increase in addition to out-of-pocket costs. This patient may be worse off after the introduction of such technological change than before its implementation.

Costs in New Healthcare Systems

Given the generation and implementation of high-cost technologies, healthcare is facing various challenges; in particular, the cost containment which is greatly connected with quality improvement, accountability, affordability, and the equitable access to care for all citizens and residents. Fuchs (1996) argues that "integrated healthcare systems" are the best way to deliver cost-effective care, physicians having "the *incentive*, the *information,* and the *infrastructure* needed to make […] decisions in a cost-effective way" (pp. 17). Fuchs contends that integrated healthcare systems cope with the following current and future major key challenges.

1. A first challenge consists in developing the physician's central position in medical decision-making, because the latter prescribes drugs, orders tests, and other procedures, admits patients to hospitals, and refers patients to specialists, among others. Through the promotion of a patient-physician relationship based on mutual recognition, accountability and acceptance of rights, patients become informed and can act cooperatively with thoughtful and careful physicians. Such measures may prevent patient moral hazard and physician malpractices.
2. With regard to the physician payment, the Fuchs's recommendation is to implement a reimbursement that promotes a good care and balances the obligations, rights, and income of physicians. Fuchs advocates capitation and adjustment for patient characteristics. This Fuchs's recommendation deserves the following details. Capitation is a system where a per capita amount is paid to a physician for each patient. Under its initial form, this amount is determined by a per member per month calculation. Clearly, under this per-member-per-month payment, the physician income depends on the *numbers of patients* to whom the healthcare service is delivered, which may lead to negligence. However, it is insightful to note that this system could give incentives to promote a *preventive* medicine since, receiving a fixed amount per patient, it is in the physician's interest to have healthy patients in order to decrease the numbers of their visits. Nevertheless, it is also worthwhile to note that a salary payment, characterized by a predetermined amount whatever the numbers of patients, may prevent carelessness and other malpractices.
3. Another Fuchs's recommendation relates to copayments at the time of use—except for the poor—and to a basic healthcare plan, plus a variety of options corresponding to services (choice of health providers, access to new experimental technologies, as well as to the older technologies not included in the basic plan) desired and paid by the patient. Fuchs specifies that such options are not alternative insurance plans, and do not create a separate plan between the poor and the rich.
4. Technologies used for diagnosis and treatment are characterized by expensive costs so that practicing in a costeffective way is not an easy task for physicians. Furthermore, this difficulty is increased by the great uncertainty of such technologies with regard to their merits. Consequently, Fuchs suggests the creation of a Center for Technology Assessment (CTA) to provide people and policy makers with independent, timely, and comprehensive assessments of technological impacts on society. Two main functions are recommended for such a center. First, it would develop knowledge about the costeffectiveness of medical technology, provide health professionals with information to measure, improve their clinical practices, and select the medical services that should be included in the basic integrated system. Second, it would give legitimacy for the costeffective medical practice. Recently, an international CTA emerged. It is worthwhile to note that, besides an "appropriate technology" (adaptable to local needs) for diagnosis and treatment, integrated systems do not forget the holistic medicines which are more time intensive than capital intensive (seeming thus favoring the salary payment than the capitation payment) and which also further affordability and accountability.

Healthcare Information Systems faced to Costs and Technological Change

New healthcare organizations meet the increasing demands of healthcare access and quality improvement as well as reduced costs. To reach such challenges and mitigate informational inadequacies (asymmetric, imperfect, and incomplete information), some new regulations and information systems are required. Various telecommunication networks—in particular, e-healthcare systems—promote patient-centric organizations and institutions which encourage patients' involvement in their own care, while providing relevant information for practitioners. Through such information systems, one can witness to the containment of various transaction costs such as: (1) the "search costs" and "communication costs" between the patients and the practitioners; (2) the "decision costs" since healthcare providers know immediately the patients' characteristics through the electronic medical records and the integrated information management systems; and (3) the "enforcement costs," since the practitioners can quickly prescribe a treatment and/or order laboratory tests.

On another level, such patient-centric healthcare systems can also cut the costs of the preventive, diagnostic, and curative medicines. (1) Indeed, by promoting investment in *preventive medicine*, such healthcare information systems may bring down medical costs, such a prevention averting the risk of disease. (2) With regard to the *diagnostic medicine*, electronic medical records and integrated information management systems, by reducing the spatial and temporal costs of a face-to-face care, may contribute also to rein in medical costs in the context of quality and safety clinical care. Diagnostics carried out immediately under full information about the patient history are major factors to improve quickly the quality of care and thus reduce medical costs. (3) Concerning the *curative medicine*, information systems may quickly order medications, laboratory work, and new diagnostic tests if necessary. Finally, information systems provide also an increased access to care for the poor and patients living in distant rural areas.

Through acts—such as the Health Insurance, Portability, and Accountability Act (HIPAA) enacted by the United States or directives (e.g., European Union Directives)—other various decisions have been recently taken for regulate the healthcare policy (for a presentation of the HIPAA's key elements, see Fadlalla & Wickramasinghe, 2004; Wickramasinghe, 2007). Recently, in the specific context of the Healthcare Information System (HCIS), Wickramasinghe (2007, pp. 8–9) contends that "information flowing within the HCIS and between […] key players must exhibit the attributes and the dimensions of the *information integrity* construct." Information integrity "should display the attributes of accuracy, consistency, and reliability of the content and processes as well as the dimensions of usefulness, completeness, manipulability, and usability" (Wickramasinghe, 2007, pp. 9).

FUTURE TRENDS

This article approaches some of the issues that must be considered when analyzing the relationships between healthcare costs and the introduction and implementation of new technologies. Efficient solutions exist to reduce transaction costs, defined as coordination and informational costs. However, our account of this topic is, of course, incomplete, and many questions still remain unanswered. Future studies could pay more attention on some important issues. (1) The relationship between the new technologies (biotechnologies and biomedical engineering, among others), and the cost containment needs to be scrutinized. (2) The relationship between such technologies and the insurance question requires further research. (3) With regard to e-health systems: *(i)* the study of their strengths and weaknesses on healthcare quality improvement requires more theoretical and empirical attention. The ways to increase social responsibilities and community values to increase such quality is a valued perspective; *(ii)* their impact on practitioner payments and health insurance need also further studies; *(iii)* however, in some contexts, such information systems must be particularly vigilant on the questions relating to prevention programs (early detection), holistic options, and affordability. Keeping in mind that all the countries do not have the same medical needs, nor the same resources and medical practices, creative solutions to promote healthcare quality improvement and cost containment must be continuously undertaken. The development of health prevention and cooperation between various organizations offer promising results.

CONCLUSION

The critical issues discussed in this article summarize many challenges to healthcare cost containment. This article studies connections between instruments and activities, the introduction and implementation of new technologies affecting healthcare processes. Nowadays, achieving more quality improvement and cost containment constitutes a challenge for healthcare systems. Thus, new e-healthcare systems succeed in internalizing many externalities, in particular those concerning the myriad of informational issues. How such systems will evolve and what consequences they have remain challenges for further research.

REFERENCES

Akerlof, G. (1970). The market for lemons: Qualitative uncertainty and the market mechanism. *Quaterly Journal of Economics*, *84*, 488–500.

Arrow, K.J. (1963). Uncertainty and the welfare economics of medical care. *American Economic Review*, *53*, 941–973.

Dranove, D. (2000). *The economic evolution of American healthcare. From Marcus Welby to managed care*. Princeton University Press.

Ellis, R.P., & McGuire, T.G. (1993). Supply-side and demand-side cost sharing in healthcare. *Journal of Economic Perspectives*, *7*(4), 135–151.

Evans, R. (1974). Supplier-induced demand: Some empirical evidence and implications. In M. Perlman (Ed.), *The Economics of Health and Medical Care*. London: MacMillan.

Fadlalla, A., & Wickramasinghe, N. (2004). An integrative framework for HIPAA-compliant I*IQ healthcare information systems. *International Journal of Healthcare Quality Insurance*, *17*(1), 65–74.

Folland, S., Goodman, A.C., & Stano, M. (2001). *The economics of health and healthcare*. New Jersey: Prentice Hall.

Fuchs, V. R. (1996). Economics, values, and healthcare reform. *American Economic Review*, *86*(1), 1–24.

Goodeeris, J.H. (1984a). Medical insurance, technological change, and welfare. *Economic Inquiry*, *22*, 56–67.

Goodeeris, J.H. (1984b). Insurance and incentives for innovation in medical care. *Southern Economic Journal*, *51*(2), 530–539.

Holmström, B. (1979). Moral hazard in teams. *Bell Journal of Economics*, *13*, 324–340.

Keeler, E.B., Brook, R.H., Goldberg, G.A., Kamberg, C.J., & Newhouse, J.P. (1985). How free care reduced hypertension in the health insurance experiment. *Journal of the American Medical Association*, *254*(14), 1926–1931.

Lurie, N., Kamberg, C.J., Brook, R.H., Keeler, E.B., & Newhouse, J.P. (1989). How free care improved vision in the health insurance experiment. *American Journal of Public Health*, *79*(5), 640–642. (Erratum in *American Journal of Public Health*, *79*(12), 1677).

Milgrom, P., & Roberts, J. (1992). *Economics, organization and management*. Prentice-Hall International Editions.

Newhouse, J.P. (1988). Has the erosion of the medical marketplace ended? *Journal of Health Politics, Policy and Law*, *13*, 263–277.

Pauly, M.V. (1994). Editorial: A re-examination of the meaning and importance of supplier-induced demand. *Journal of Health Economics*, *13*, 369–372.

Peden, E. A., & Freeland, M.S. (1998). Insurance effects on U.S. medical spending (1960–1993). *Health Economics*, *7*, 671–687.

Reinhardt, U.E. (1989). Economists in healthcare: Saviors, or elephants in a porcelain shop? *American Economic Review*, *79*, 337–342.

Rice, T.H., & Labelle, R.J. (1989). Do physicians induce demand for medical service? *Journal of Health Politics, Policy and Law*, *14*, 587–600.

Robinson, J., & Luft, H. (1985). The impact of hospital market structure on patient volume, average length of stay, and the cost of care. *Journal of Health Economics*, *4*, 333–356.

Roemer, M.I. (1961). Bed supply and hospital utilization: A national experiment. *Hospitals, The Journal of the American Hospital Association*, *35*, 36–42.

Rothschild, M., & Stiglitz, J.E. (1976). Equilibrium in competitive insurance markets. *Quaterly Journal of Economics*, *90*(4), 630–649.

Scitovsky, A.A. (1985). Changes in the costs of treatment of selected illnesses, 1971–1981. *Medical Care*, *23*, 1345–1357.

Wickramasinghe, N. (2007). Patient centric healthcare information systems in the U.S. *Mimeography*. Center for the Management Medical Technology, Chicago.

Williamson, O.E. (1979). Transaction-cost economics: The governance of contractual relations. *Journal of Law and Economics*, *22*, 233–261.

KEY TERMS

Agency Relationship: Situation in which one party relies on another party to make decisions.

Copayment: The fixed partial payment that a subscriber to a medical insurance must pay for the use of a medical service covered by his/her insurance company.

Curative Medicine: Actions and treatments correcting a harmful or troublesome situation.

Diagnostic Medicine: Act or process of identifying the cause and nature of the illness by its signs and symptoms.

Moral Hazard: In economic theory, moral hazard refers to the postcontractual opportunism that results from an asymmetric information. It occurs when the actions contractually required are not freely observable; for example, a party takes an inappropriate action or decision that can affect the outcomes of the contract. In insurance, moral hazard refers to the tendency of people to change their behavior when they are insured; it does not induce the insured to protect him/herself from the risk.

Preventive Medicine: Act of protecting, promoting, maintaining health, and preventing disease.

Transaction Costs: Costs incurred in making an transaction. They include costs of coordinating (search, communication, decision, enforcement costs) and costs of motivating (opportunity costs).

C

Cost-Sensitive Classification for Medical Diagnosis

Gerald Schaefer
Aston University, UK

Tomoharu Nakashima
Osaka Prefecture University, Japan

Yasuyuki Yokota
Osaka Prefecture University, Japan

INTRODUCTION

Medical diagnosis is often regarded as a pattern classification problem: based on a certain input, the task is to assign it to one of a set of classes. Often the number of classes is two: for example, malignant vs. benign, disease vs. no disease. The task of a pattern classification system is to assign as many input samples as possible to the correct class whereas the behaviour of the classifier is often optimised through the learning of some ground truth data. Conventional classifiers treat each sample of this learning set equally, yet in medical diagnosis this is often not desirable as different classes are associated with different costs. While the misdiagnosis of a malignant case as being benign can be very costly (e.g., when the time for effective treatment has passed), mistaking a benign case as malignant (though of course it should be avoided) will involve relatively lower costs (such as some further tests).

In this article, we present a cost-sensitive approach to medical diagnosis based on fuzzy rule-based classification (Schaefer, Nakashima, Yokota, & Ishibuchi, 2007). While fuzzy rule-based systems have been mainly employed for control problems (Lee, 1990) more recently they have also been applied to pattern classification problems (Ishibuchi & Nakashima, 1999; Nozaki, Ishibuchi, & Tanaka, 1996). We modify a fuzzy rule-based classifier to incorporate the concept of weight which can be considered as the cost of an input pattern being misclassified. The pattern classification problem is thus reformulated as a cost minimisation problem. Based on experimental results on the Wisconsin breast cancer dataset, we demonstrate the efficacy of our approach. We also show that the application of a learning algorithm can further improve the classification performance of our classifier.

FUZZY RULE-BASED CLASSIFICATION

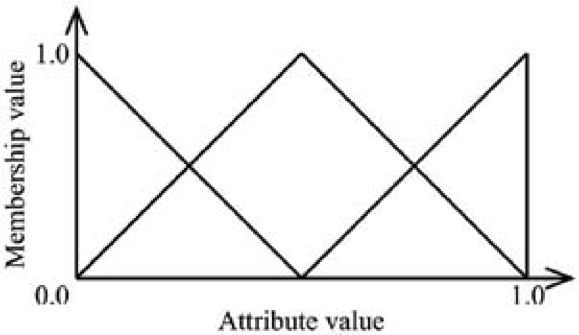
Let us assume that our pattern classification problem is an n-dimensional problem with M classes (in medical diagnosis M is typically 2) and m given training patterns $\mathbf{x}_p = (x_{p1}, x_{p2},...,x_{pn})$, $p = 1,2,...,m$. Without loss of generality, we assume each attribute of the given training patterns to be normalised into the unit interval [0,1]; that is, the pattern space is an n-dimensional unit hypercube $[0,1]^n$. In this study we use fuzzy if-then rules of the following type as a base of our fuzzy rule-based classification systems:

$$\text{Rule } R_j: \text{ If } x_1 \text{ is } A_{j1} \text{ and } \ldots \text{ and } x_n \text{ is } A_{jn}$$
$$\text{then Class } C_j \text{ with } CF_j \qquad j = 1,2,\ldots,N \qquad (1)$$

where R_j is the label of the j-th fuzzy if-then rule, $A_{j1},...,A_{jn}$ are antecedent fuzzy sets on the unit interval [0,1], C_j is the consequent class (i.e., one of the M given classes), and CF_j is the grade of certainty of the fuzzy if-then rule R_j. As antecedent fuzzy sets we use triangular fuzzy sets as in Figure 1 where we show a partition of the unit interval into a number of fuzzy sets.

Our fuzzy classification system consists of N fuzzy if-then rules each of which has a form as in Equation (1). There are two steps in the generation of the rules: specification of antecedent part, and determination of consequent class C_j and grade of certainty CF_j. Once the antecedent part of a rule is specified the consequent part (i.e., consequent class and the grade of certainty) is determined from a set of given training patterns (Ishibuchi, Nozaki, & Tanaka, 1992). In Ishibuchi and Nakashima (2001), it is shown that the use of the grade of certainty in fuzzy if-then rules allows us to generate comprehensible fuzzy rule-based classification systems with high classification performance.

Figure 1. Triangular membership function



Fuzzy Rule Generation

Let us assume that m training patterns $\mathbf{x}_p = (x_{p1},...,x_{pn})$, $p = 1,...,m$ are given for an n-dimensional C-class pattern classification problem. The consequent class C_j and the grade of certainty CF_j of the if-then rule are determined in the following manner:

1. Calculate $\beta_{Class\ h}(j)$ for Class h as:

$$\beta_{Classh\ h}(j) = \sum_{\mathbf{x}_p \in Class\ h} \mu_j(\mathbf{x}_p) \tag{2}$$

where

$$\mu_j(\mathbf{x}_p) = \mu_{j1}(x_{p1}) \cdot ... \mu_{jn}(x_{pn}) \tag{3}$$

and $\mu_{jn}(\cdot)$ is the membership function of the fuzzy set A_{jn}. In this article, we use triangular fuzzy sets as in Figure 1.

2. Find Class $\hat{h}$ that has the maximum value of $\beta_{Class\ h}(j)$:

$$\beta_{Class\ \hat{h}}(j) = \max_{1 \le k \le C}\{\beta_{Class\ k}(j)\} \tag{4}$$

If two or more classes take the maximum value, the consequent class C_j of the rule R_j cannot be determined uniquely. In this case, we specify C_j as $C_j = \{\}$. If a single class $\hat{h}$ takes the maximum value, let C_j be Class $\hat{h}$.

The grade of certainty CF_j is determined as:

$$CF_j = \frac{\beta_{Class\ \hat{h}}(j) - \bar{\beta}}{\sum_h \beta_{Class\ h}(j)} \tag{5}$$

with

$$\bar{\beta} = \frac{\sum_{h \ne \hat{h}} \beta_{Classh}(j)}{c-1} \tag{6}$$

Fuzzy Reasoning

Using the rule generation procedure outlined above we can generate N fuzzy if-then rules as in Equation (1). After both the consequent class C_j and the grade of certainty CF_j are determined for all N rules, a new pattern $\mathbf{x} = (x_1,...x_n)$ can be classified by the following procedure:

1. Calculate $\alpha_{Class\ h}(\mathbf{x})$ for Class h, j=1,...,C, as:

$$\alpha_{Class\ h}(\mathbf{x}) = \max\{\mu_j(\mathbf{x}) \times CF_j \mid C_j = h\} \tag{7}$$

2. Find Class h' that has the maximum value of $\alpha_{Class\ h}(\mathbf{x})$:

$$\alpha_{Class\ h'}(\mathbf{x}) = \max_{1 \le k \le C}\{\alpha_{Class\ k}(\mathbf{x})\} \tag{8}$$

If two or more classes take the maximum value, then the classification of **x** is rejected (i.e., **x** is left as an unclassifiable pattern), otherwise we assign **x** to Class h'.

COST-SENSITIVE FUZZY CLASSIFICATION

The standard fuzzy rule-based classifier as detailed above treats each class and hence each sample equally. In medical diagnosis however this is often not desirable. Misdiagnosing a malignant case as benign should be penalised more than diagnosing healthy patients as having a certain disease. While in the first case, the result might be that of late treatment in the best and missing of the treatable time in the worst scenario, the latter case will usually involve some further tests which should then identify the misdiagnosis. Clearly, the costs involved in the first case will exceed those of the latter.

We therefore wish to develop a classifier that incorporates this and reformulate the pattern classification problem as a cost minimisation problem. The concept of a weight is introduced for each training pattern in order to handle this situation where the weight of an input pattern can be viewed as the cost of misclassifying it. Fuzzy if-then rules are generated by considering the cost as well as the compatibility of training patterns.

In order to incorporate the concept of cost, we modify Equation 2 of the fuzzy rule generation to:

$$\beta_{Class\ h}(j) = \sum_{\mathbf{x}_p \in Class\ h} \mu_j(\mathbf{x}_p)\,\omega_p \quad (9)$$

where ω_p is the cost associated with training pattern p. A suitable overall cost function can be defined as:

$$Cost(S) = \sum_{p=1}^{m} \omega_p \cdot z_p(S), \quad (10)$$

where m is the number of training patterns, ω_p is the weight/cost of the training pattern $\mathbf{x}_p$, and $z_p(S)$ is a binary variable set according to the classification result of the training pattern $\mathbf{x}_p$ by S: $z_p(S) = 0$ if $\mathbf{x}_p$ is correctly classified by S, and $z_p(S) = 1$ otherwise (i.e., $\mathbf{x}_p$ is misclassified or rejected). We use this cost function as a performance measure as well as classification rate.

LEARNING ALGORITHM

A learning method that adjusts the grades of certainty CF_j can be employed to achieve improved classification performance (Nakashima, Yokota, Ishibuchi, & Schaefer, 2005). It is based on an error-correction learning approach where the adjustment occurs when classification of training patterns is not successful. When a training pattern is correctly classified, we do not adjust the grade of certainty. The main idea of the learning method is to adjust the degree of certainty CF_j of two fuzzy if-then rules: We decrease the degree of certainty of a fuzzy if-then rule that misclassifies a training pattern and in turn increase that of a fuzzy if-then rule that is supposed to correctly classify the training pattern.

Let us assume that we have generated fuzzy if-then rules by the rule-generation procedure detailed above. We also assume that a fuzzy if-then rule R_j misclassifies a training pattern $\mathbf{x}_p$. That is, R_j is used to classify $\mathbf{x}_p$ from Class c^* by using Equation (8) but the consequent class C_j does not agree with the true class of the training pattern $\mathbf{x}$. Let R_* be the fuzzy if-then rule that is selected by Equation (7). That is, R_* has the maximum value of $\alpha_{class\ c^*}(\mathbf{x}_p)$ among those fuzzy if-then rules with Class c^* but does not have the maximum value among all generated fuzzy if-then rules. The proposed learning method adjusts the grades of certainty of R_j and R_* as follows:

$$CF_j^{new} = CF_j^{old} - \eta \cdot \omega_p \cdot CF_j^{old} \quad (11)$$

$$CF_*^{new} = CF_*^{old} - \eta \cdot \omega_p \cdot (1 - CF_*^{old}) \quad (12)$$

where ω_p is the weight of the training pattern $\mathbf{x}_p$, and η is a positive constant value. We assume that $0 \le \eta \le 1$.

One epoch of the proposed learning method involves examining all given training patterns. Thus there will be $2m$ adjustments of fuzzy if-then rules if all m training patterns are misclassified. The learning process is summarised as follows:

1. Generate fuzzy if-then rules from m given training patterns.
2. Set K as K=1.
3. Set p as p=1.
4. Classify $\mathbf{x}_p$ by using the generated fuzzy if-then rules in Step.
5. If $\mathbf{x}_p$ is misclassified, adjust the grades of certainty using Equations (11) and (12). Otherwise no rules are adjusted.
6. If $p<m$, let $p:=p+1$ and go to Step . Otherwise go to Step.
7. If K reaches a pre-specified value, stop the learning procedure. Otherwise let $K:=K+1$ and go to Step.

Note that K in the above learning procedure corresponds to the number of epochs.

EXPERIMENTAL RESULTS

In order to evaluate our proposed cost-sensitive fuzzy classifier, we tested it on a standard medical classification dataset, namely the Wisconsin breast cancer dataset (Wolberg & Mangasarian, 1990) which comprises a collection of nine cytological attributes such as clump thickness, uniformity of cell size and shape, and so forth, for 444 benign and 239 malignant cases.

We used a standard fuzzy rule-based classifier to obtain a baseline benchmark to compare our algorithms to. We then applied our proposed cost-based classifier with two different cost setting: a ratio of 1:2 between benign and malignant cases and a ratio of 1:5 (though both are probably still conservative estimates). We also used the learning algorithms detailed above to improve upon the classification performance of our cost-based classification system. We investigate two different sets of parameters for the learning algorithm: a slower learning method with η=0.2 and *K*=50 and a faster approach with η=0.5 and *K*=20. In all experiments, we divide each attribute uniformly into three triangular fuzzy sets as shown in Figure 1. In order to arrive at statistically meaningful results, in all cases we perform 10-fold cross-validation where the patterns are split into 10 disjoint subsets and each subset is in turn used as an unseen test set while the other nine sets are used for training the classifier. We report the results in terms of sensitivity defined as:

$$SE = \frac{TP}{TP + FN}$$

and specificity defined as:

$$SP = \frac{TN}{TN + FP}$$

where *TP*, *TN*, *FP*, and *FN* correspond to true positives, true negatives, false positives, and false negatives respectively. All results are given as the average 10-CV scores for both training and unseen test data and are shown in Tables 1 and 2.

On the training data, the conventional classifier achieves a sensitivity and specificity of 97.49% and 98.65% respectively, which corresponds to six false negatives and six false positives. Our cost-based ap-

Table 1. 10-CV results on breast cancer training data

	Classifier	tot. cost	SE (%)	SP (%)	TP	FN	FP	TN
1:2	conventional	18	97.49	98.65	233	6	6	438
	cost-based	15	98.33	98.42	235	4	7	437
	cost+learning η=0.2,*K*=50	2	100.00	99.55	239	0	2	442
	cost+learning η=0.5,*K*=20	2	100.00	99.55	239	0	2	443
1:5	conventional	36	97.49	98.65	233	6	6	438
	cost-based	21	99.16	97.52	237	2	11	433
	cost+learning η=0.2,*K*=50	2	100.00	99.55	239	0	2	442
	cost+learning η=0.5,*K*=20	2	100.00	99.55	239	0	2	443

Table 2. 10-CV results on breast cancer test data

	Classifier	tot. cost	SE (%)	SP (%)	TP	FN	FP	TN
1:2	conventional	39	94.14	97.52	225	14	11	433
	cost-based	37	95.40	96.66	228	11	15	429
	cost+learning η=0.2,*K*=50	37	95.40	96.66	228	11	15	429
	cost+learning η=0.5,*K*=20	37	95.40	96.66	228	11	15	429
1:5	conventional	81	94.14	97.52	225	14	11	433
	cost-based	56	96.65	96.40	231	8	16	428
	cost+learning η=0.2,*K*=50	56	96.65	96.40	231	8	16	428
	cost+learning η=0.5,*K*=20	56	96.65	96.40	231	8	16	428

proach improves upon this by correctly identifying two more malignant cases (while increasing the number of false positives by 1). A further dramatic improvement is observed after we apply the learning algorithm achieving a sensitivity of 100% with a specificity of 99.55%.

On the test data the difference between the algorithms is still significant though not as pronounced as for the training data. Here, three more malignant cases are identified for a cost setting of 1:2 while six more cancer patients are detected with a cost setting of 1:5. The effect of the misclassification costs on the performance is hence readily observable here; also in all cases the cost-based variations produce lower overall costs compared to the conventional approach.

FUTURE TRENDS

As has been shown in this article, the application of costs and an associated cost-sensitive classifier can prove very useful for medical diagnosis. The main problem here, however, is that in most cases the costs are only estimated and hence do not necessarily represent the true costs. Consequently, the derived classifier might not perform as well as it could. A different approach would, hence, be to derive a classifier that does not take into account the costs directly but rather relies on an ordering of importance between the different classes (Nakashima, Yokota., Schaefer, & Ishibuchi, 2007).

Also, the fuzzy rule base derived in our work contains some redundancies and is hence not as efficient and compact as possible. To arrive at a smaller but equivalent rule base, rule pruning methods can be applied. A smaller rule base also has the advantage of better interpretability of the derived rules.

CONCLUSION

In medical diagnostic classification systems, classification performance is not always the only indicator for assessing classifiers. Rather, misclassification costs should be taken into account as well, as usually misclassifying a malignant case will prove much more costly than misclassifying a benign one. In this article, we have applied a cost-sensitive fuzzy rule-based classifier, which emphasises the importance of those classes which have high misclassification costs, to medical diagnostic classification datasets. We have also applied a learning algorithm to further boost the classification results. Experimental results have shown that our cost-based approaches perform better than conventional classifiers under the assumed conditions.

REFERENCES

Ishibuchi, H., & Nakashima, T. (1999). Performance evaluation of fuzzy classifier systems for multi-dimensional pattern classification problems. *IEEE Trans. Systems, Man and Cybernetics (Pt. B): Cybernetics, 29*, 601-618.

Ishibuchi, H., & Nakashima, T. (2001). Effect of rule weights in fuzzy rule-based classification systems. *IEEE Trans. Fuzzy Systems, 9*(4), 506-515.

Ishibuchi, H., Nozaki, K., & Tanaka, H. (1992). Distributed representation of fuzzy rules and its application to pattern classification. *Fuzzy Sets and Systems, 52*(1), 21-32.

Lee, C.C. (1990). Fuzzy logic in control systems: Fuzzy logic controller (Pt. I and Pt. II). *IEEE Trans. Systems, Man and Cybernetics, 20,* 404-435.

Nakashima, T., Yokota, Y., Ishibuchi, H., & Schaefer, G. (2005). Learning fuzzy if-then rules for pattern classification with weighted training patterns. In *4th Conference of the European Society for Fuzzy Logic and Technology* (pp. 1064-1069).

Nakashima, T., Yokota, Y., Schaefer, G., & Ishibuchi, H. (2007). Introducing class-based classification priority in fuzzy rule-based classification systems. *IEEE International Conference on Fuzzy Systems*, 1757-1762.

Nozaki, K., Ishibuchi, H., & Tanaka, H. (1996). Adaptive fuzzy rule-based classification systems. *IEEE Trans. Fuzzy Systems, 4*(3), 238-250.

Schaefer, G., Nakashima, T., Yokota, Y., & Ishibuchi, H. (2007). Cost-sensitive fuzzy classification for medical diagnosis. *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, 312-316.

Wolberg, W.H., & Mangasarian, O.L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. In *Proceedings of the National Academy of Sciences, 87,* 9193-9196.

KEY TERMS

Cross Validation: A standard method of evaluating classifiers where data with known classes is divided into disjoint sets and the classifier tested on various combinations of these training sets.

Fuzzy Logic: A form of logic where variables can take on variable degrees of truth.

Fuzzy Set: An extension of the classical set whose memberships have degrees of membership.

Membership Function: A function that describes the degree of an element's membership in a fuzzy set.

Pattern Classification: Automatic transformation of input data into categories (classes).

Rule-Based Classifier: A pattern classification system in which the classification is expressed as a set of rules.

Creating and Marketing a Dental Supply Online Store

David Gadish
California State University, USA

INTRODUCTION

The dental supply company is family owned and operated. It was founded in the 1940s and currently serves approximately 1,000 dental offices and offers over 16,000 different products from over 200 manufacturers. Management did not consider the importance of information technology solutions that other firms have implemented to increase efficiency and cut costs. A decision to modernize came in 2003 as revenues started falling and existing clients started ordering dental supplies from competitors. This analysis covers the first 2 years of marketing from November 2004 to November 2006. The process was preceded with a 2-month campaign planning phase.

Understanding the Dental Supply Industry

The North American market size in 2005 was estimated to be $5.6 billion (Patterson, 2006). Two main players in the market are Patterson with 32% and Sullivan-Schein with 30% market share. Market growth is estimated to be 7% to 9% annually.

The Patterson Dental Company provides a range of consumable dental products, clinical and laboratory equipment, and value-added services to dentists, dental laboratories, institutions, and other health care providers. Patterson Dental has the largest direct sales force in the industry, totaling nearly 1,300 sales representatives and equipment/software specialists. Projected revenues for the fiscal year 2005 were 1.8 billion dollars (32% of the North American market share) (Patterson, 2006). Sullivan-Schein Dental has an extensive line of dental, medical, and pharmaceutical products (Sullivan-Schein, 2006), the company holds about 30% of the North American market in fiscal year 2005 (Patterson, 2006). Benco Dental serves 20,000 customers in 18 states and employs about 950 associates (Benco, 2006).

Pearson Dental Supplies offers dental supplies and dentist office equipment. The company was founded in 1945 and its supply catalog carries over 65,000 products, making it one of the big players in the dental supply industry (Pearson, 2006).

The other players in the Dental Supply Industry include but are not limited to AccuBite Dental Supply, Atlanta Dental, Applied Dental, Burkhart Dental, Conger Dental Supply Company, Darby Dental Supply, Discus Dental, and Eastern Dental Supply.

The North American dental market consists of about 156,000 U.S. dentists and around 18,000 Canadian dentists (Pearson, 2006). 65%-70% are sole practitioners. There are over 135,000 dental practices. The average revenue per dentist stands at $550,000 per annum. Dentists spend $0.05 – $0.07 revenue per dollar on consumables supplies, which means $25,000 to $35,000 of revenues per annum.

The international market has enormous possibilities for U.S. exports. The European Union will continue to be one the largest export markets for the U.S., but countries in Central and South America and India are becoming larger buyers of high quality dental supply.

The U.S. has historically accounted for roughly 50% of the global market for dental equipment and supplies. In 1998, the total export of dental products totaled $633 million. In 1998, exports to Europe rose to $277 million, which represented 44% of total exports of dental supplies. As Japan and China become world powers, their need for dental hygiene products increases. As of 1998, U.S. sales reached $28 million, which represented a 40% increase since 1996. This is attributed to economic growth, higher income levels, increased access to dental clinics, and great awareness of dental hygiene (Palmer, 2002).

The Internet, the World Wide Web, and other developments of the information revolution will redefine patient care, referral relationships, practice management, quality, professional organizations, and competition (Bauer & Brown, 2001).

THE DENTAL SUPPLY E-COMMERCE SOLUTION

The Web offers the advantages of both centralization of information and coordination (Marks, 2004). To survive, the company needed to increase revenues and reduce operating costs. They chose to accomplish this by automating business processes through the implementation of their online solution. The Web site provides customers with information online. It allows dentists to place orders online. It also allows prospective dentists to reach them. The company will become more accessible to their clients and prospects (Goff, 2002). It will also be able to further build on its brand.

The company uses their Web site to further inform its customers of new dental supplies and sell dental supplies to its customers. The company's Web site integrates three areas of commerce:

- The Products system includes product search and an online catalog system
- The Customers system includes a customer accounts systems, authentication system and links to the order and product systems
- The Orders system includes customer orders and delivery systems

Databases, Web servers, Web applications, and other computer technology were used to get the Web site up and running. Data about dental products was loaded from the company's existing ERP system. Product images were requested from a few hundred dental product manufacturers who's products the company resells, and were then incorporated into the e-commerce solution. The process is detailed in the following sections.

Planning of Dental Supply E-Commerce Infrastructure

In creating the e-commerce solution for ENG Dental Supply, the Systems Development Life Cycle process was utilized. It consists of seven phases. The planning phase covered project scope and budget. The analysis phase included research of the competition, analysis of employees and client requirements, research of ISPs, and research of e-commerce packages. The design phase included designing processes to add/remove/modify products and planning marketing strategies. The system build phase includes the purchasing and configuring test computer, purchasing the e-commerce package, and development and testing of an e-commerce Web site. The data build phase involved many activities including: extraction of product data from the DMS system, identification of products to list online, cleanup of product data, classification of products, data load, collection of product images, population of images into the e-commerce Web site, and finally testing the data in the e-commerce Web site. The deployment phase consisted of: creation of training materials, training employees, moving the site from test to production environment, and training clients online. The final phase is operations, which included activities; updating prices, adding/removing/modifying products, and marketing and supporting the e-commerce Web site. Once these activities were identified, a preliminary project budget was established (see Figure 1).

Determine Project Scope

ENG needed to catch up with their competitors almost immediately if they wanted to remain a viable dental supply company. In order to do so, ENG needed to provide its customers with an online product catalog that would have electronic commerce capabilities so the customers could compare products and place orders.

Determine Budget

The client allocated $45,000 for the initial implementation, and so the question facing the team was how to implement the maximum possible for the tight budget.

Figure 1. Dental supply e-commerce planning phase

To achieve that, the activities were prioritized and implemented selectivity.

Analysis of Dental Supply E-Commerce

The e-commerce site needed to be easy to use and maintain. An analysis study of the company's business processes was performed in order to identify ENG's requirements. Information about employees and their duties was collected, and a course of actions was formulated based on the analysis of the findings (see Figure 2).

Research Competition

Portals differed in many characteristics, such as the number of services, product pricing, discussion forum activity, navigability, reaction time in response to questions, and site responsiveness (Schleyer & Spallek, 2002). The online presence of the dental supply competitors including their e-commerce capabilities was analyzed. These companies included: Patterson Dental Company, Sullivan-Schein Company, Burkhart Dental, Inc., Darby Dental Supply, and Benco Dental, Inc.

Analyze Employees and Client Requirements

Meetings with management covered the employee and client requirements. The customer service department and the outside sales people were then utilized to determine their requirements for such a system. They were asked to review competitors systems, and recommend ways in which the ENG System could be designed to achieve their needs and be better then the competition.

Research ISPs

The information of the company's Internet Service Provider (ISP) was gathered. The company hosting the Web site relied on older technology and the ISP features did not allow for the usage of dynamic information. Several ISPs were studied for reliability, connectivity, and service to support the new Web site. A company by the name of 1and1 met all the requirements for the solution.

Research E-Commerce

Several e-commerce packages were tested before making a final purchasing decision. X-cart Gold was selected to drive the new e-commerce site. The price was also very competitive compared to other vendors. The main reasons why this package was chosen are shown in Table 1.

Design of the Dental Supply E-Commerce Web Site

The design phase consisted of designing an e-commerce site look and functionality, creation of the process to update prices, creation of the process to add/remove/modify products, and the planning for marketing campaigns (see Figure 3).

Design E-Commerce Site Look/Functionality

Two main actions were taken with respect to ENG's Web site, to improve its performance and usability:

Figure 2. Dental supply e-commerce analysis phase

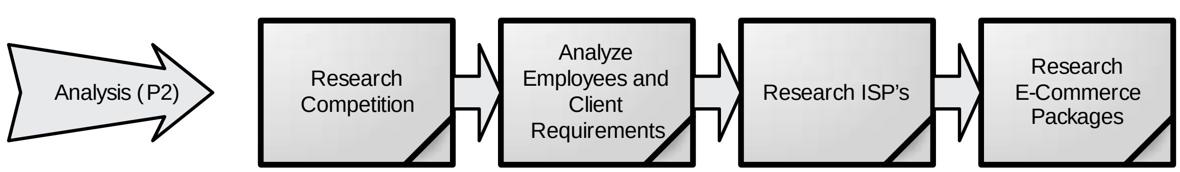


Table 1.

Free customer support for X-cart customers. Whereas most vendors provide technical support for their products at a cost, X-cart has free customer support with the purchase of the software.	Personal order history log gives ENG customers the ability to view how much and what they have ordered in the past.
Ability to print transactions.	Clients can search products by title, description and category.
The option to use discount coupon codes and gift certificates.	Full inventory control.
"Out of Stock" notifications.	Quantity discounts.
Retail and wholesale pricing.	Allow customers to choose delivery methods.
Ability to process credit cards using different known transaction companies.	Export sales and customer data for use in a spreadsheet.
Printable shipping labels.	Personalized greetings of regular customers.
Real-time order tracking for registered customers.	Encryption protection for secure transaction and customer data transmission.

Figure 3. Dental supply e-commerce design phase

- Minimize dynamic content; with careful performance analysis, it was determined that the usage of dynamic content on the site would decrease server performance. Budget constraints impeded the company to afford a high-end dedicated hosting server. Therefore, in order to improve performance on the Web site, dynamic content use would be minimized as much as possible.
- Increase the appeal of ENG's Web site. The main page of every Web site is critical in capturing the visitor's attention. Generally only 5 seconds are required to captivate the audience's attention. If the content does not convey the right message, the potential customer may be lost. The company's original Web site did not attract visitors.

The site was not sophisticated enough and e-commerce was nonexistent. A survey of 250 clients was performed in order to determine what needed to be changed. The results were very clear. The recipients agreed that the Web site needed to be thoroughly revised. Hence, the Web site was redesigned to maximize its appeal and functionality.

Create Process to Update Prices

Product prices were updated daily on the DMS environment. This was a manual process based on market prices of products offered by the competition. A new e-commerce solution meant that prices would need to be updated twice, creating possible inconsistencies between online and off-line ordering prices due to human data entry errors. To resolve this, a batch process would be created to copy all price changes from the DMS to a Microsoft Access document twice each month. This document would be manually imported into the new e-commerce environment, then updating the prices quickly and consistently.

Create Process to Add/Remove/Modify Products

A list of products added, removed or modified would be downloaded from the DMS environment twice per month. The changes to product descriptions, added products or removal records would then need to be manually recorded into the system by an operator.

Plan Marketing Campaigns

A plan and a schedule were created outlining the traditional and online marketing efforts that would take place to promote the e-commerce Web site. Some activities were planned for the prelaunch period, others for the initial operational period and for ongoing activities.

Build of the Dental Supply E-Commerce System

E-commerce software was purchased, customized, and tested (see Figure 4).

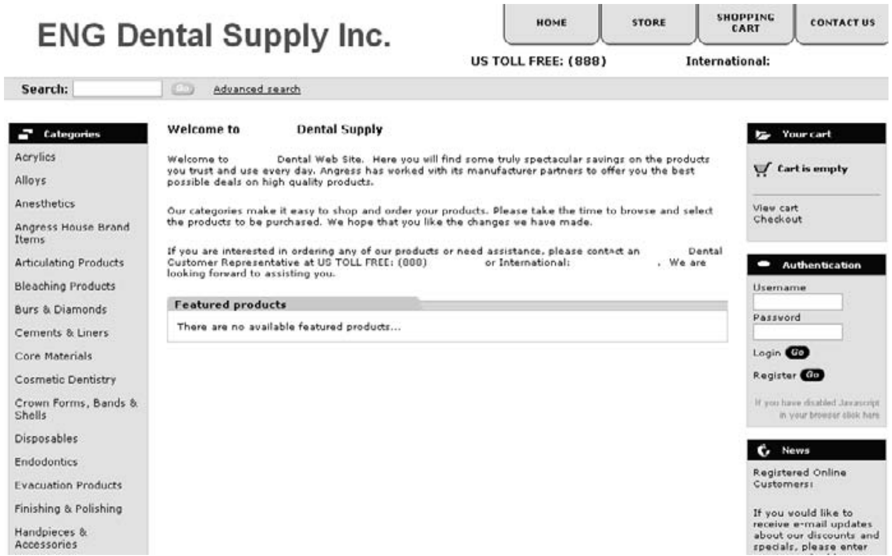
Purchase/Configure Test Computer

A dedicated computer was purchased in order to perform testing of the Web site by the company's staff. The computer was setup as a dedicated machine to perform intensive tests and identify any possible design errors. The investment of this system was a few hundred dollars.

Figure 4. Dental supply e-commerce system build phase

Figure 5. Screenshot of ENG's homepage

Figure 6. Screenshot of ENG e-commerce storefront page



Purchase E-Commerce Package

The software package chosen for the design and implementation of the e-commerce site is called X-cart Gold and was purchased from X-cart solutions (X-Cart). The basic package offered system capabilities such as querying, ordering and processing.

Create E-Commerce Web Site

The graphical user interface was customized for a more aesthetically appealing look. Features such as product search and thumbnail display were added (see Figure 5 and Figure 6).

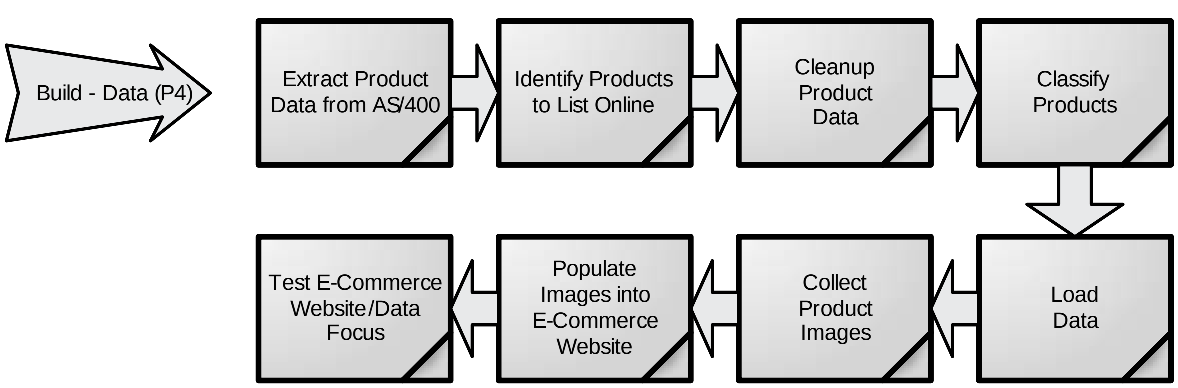
Test E-Commerce Web Site

A limited data set of product records and clients was entered into the system. Testing was performed to ensure online ordering, transaction recording, and reporting of other key functions worked flawlessly.

Build of the Dental Supply E-Commerce Data

The current database contained in the DMS was updated with detailed product descriptions and placed into a hierarchy in accordance with industry norms (see Figure 7).

Figure 7. ENG dental supply e-commerce data build phase



Extract Product Data from DMS

A script to extract all DMS data was created, tested, and executed. Data was extracted into an MS Access file.

Identify Products to List Online

The organization's purchasing manager reviewed each product item in the Microsoft Access file. Over 5,000 products were identified as being discontinued or no longer sold by the company. Records were uploaded into the e-commerce package and tested to ensure the system reflected the appropriate values.

Product Data Cleanup

Focus was placed into making sure that all products had full description including prices and tags. The data was extracted into Microsoft Excel.

Classify Products

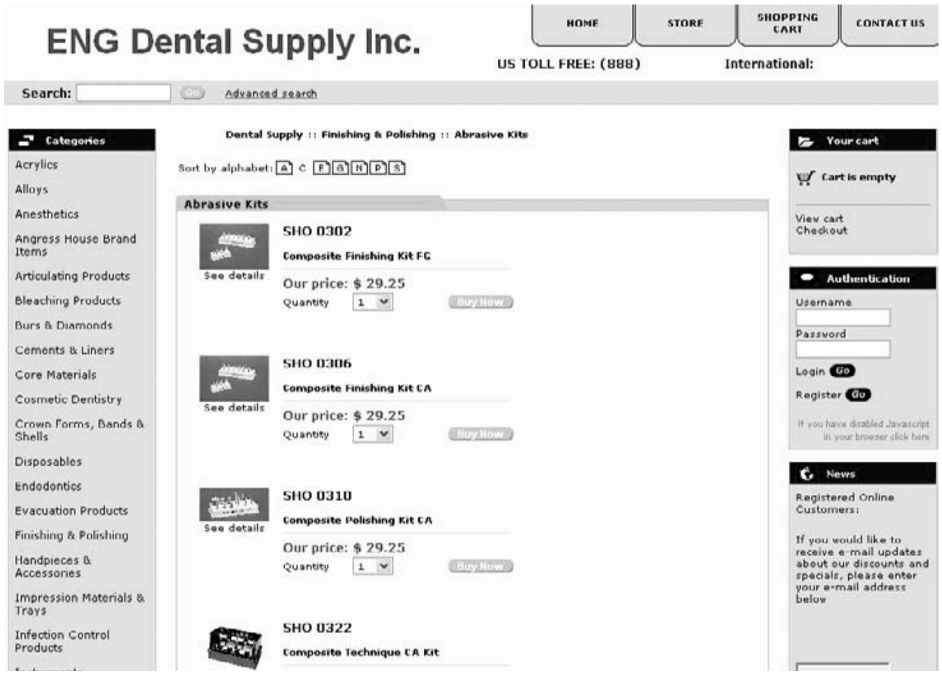
The DMS environment used only by ENG staff did not classify products in categories. Dental offices would require these to facilitate a more user friendly search. A list of ENG product categories was created by analyzing the categories of products available on competitors' Web sites (see Figure 8 and Figure 9). These categories

Figure 8. ENG's product categories

Acrylics	Cosmetic Dentistry	Infection Control of Products	Retration Materials
Alloys	Crown Forms, Bands and Shells	Instruments	Rubber Dam and Accessories
Anesthetics	Disposables	Matrix Products	Small Equipment
Articulating Products	Endodontics	Miscellaneous	Surgical Products
Bleaching Products	Evacuation Products	Orthodontics	Sutures and Suture Needles
Burs and Diamonds	Finishing and Polishing	Parts	Toys
Cements and Liners	Handpieces and Accessories	Pins and Posts	Waxes
Core Materials	Impression Materials and Trays	Preventative and Prophy Items	X-Ray Products

Figure 9. Screenshot of ENG's product categories

Figure 10. Screenshot of ENG's product listings



were subdivided into 638 subcategories to allow users to further refine their search of dental products.

Load Data

The e-commerce Web site uses MySQL to store the extracted records provided by the vendor in charge of retrieving the mainframe's information. The MS Access data was exported into a common separated value (CSV) file that would be recognized by the e-commerce package. These records were then carefully imported into MySQL to ensure data integrity (see Figure 10).

Collect Product Images

The image collection process caused major time delays due to difficulties entailed in obtaining the images from product supplier and manufacturers. ENG's suppliers and manufacturers were contacted in order to provide their images for the e-commerce environment. Manufacturers provided diskettes or CDs with the product data. These images were converted to low-resolution format using Advanced JPEG compression version 4.8 software. The images were named according to the ENG product codes. Images are in the JPG format, which is a standard format supported by the X-cart package.

Populate Images into E-Commerce Web Site

This manual process involved importing one image for the directory of images into the corresponding data record in the e-commerce environment (see Figure 11).

Test E-Commerce Web Site/Data Focus

Testing was performed for a period of one week. All departments of the organization were involved in the thorough testing of the new site. Feedback was provided for minor adjustments of the e-commerce environment. A full time person was assigned to search for the most popular products to ensure their descriptions would help locating them online. A second week of testing involved ten ENG clients. They were provided with written instructions and telephone support to help them register, and use the system to purchase dental products online.

Deployment of the Dental Supply E-Commerce Web Site

Deployment consisted of creating training materials, training employees, moving the Web site to the produc-

Figure 11. Screenshot of a dental product image

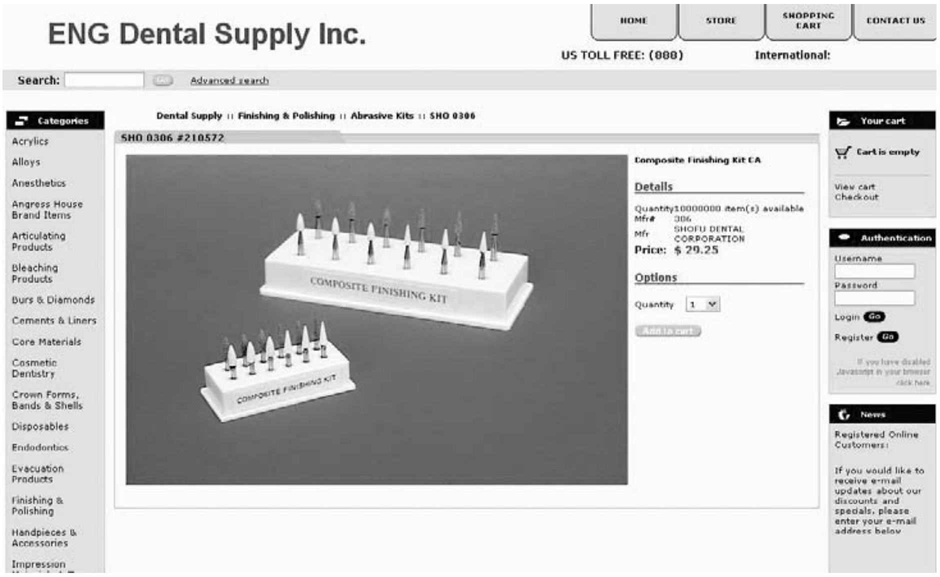


Figure 12. ENG dental supply deployment phase

tion environment, and then training ENG's customers (see Figure 12).

Create Training Materials

Two training programs were offered to ENG: one for its employees and another for its clients. Separate training materials were prepared to train customer service staff and employees supporting the system. The documentation included:

A document specifying the system's different features and capabilities; product search features, registering to the Web site, and so forth.

A document containing an overview of the architecture of the system, instructions of how to maintain and add basic features to the system.

A document containing a site map and information regarding product so as to assist customer service staff in supporting customers remotely.

Train Employees

The ENG customer service and sales department employees were trained on registering to the system, searching for products, reviewing transactions, and extracting transactions from the DMS environment for order fulfillment.

Move Site from Test to Production Environment

The new Web site had to be uploaded and promoted from testing to production once the required tests were performed.

Figure 13. ENG dental supply e-commerce operations phase

Train Clients Online

Two main options for client training support were offered:

- Phone support by the customer service department staff for those clients encountering difficulties ordering or searching products through the system.
- Personal visit by sales representatives - in extreme cases where phone support didn't resolve the clients' problems or for VIP customers.

Customers were trained to register online, to search for products, to order products, to change orders and to review the order history.

Operations of the Dental Supply E-Commerce Web Site

Ongoing operations of the e-commerce environment include price updates, adding, removing or modifying product information, marketing the e-commerce Web site, and supporting the site and the data (see Figure 13).

Update Prices

The customer service support department employees were trained in the process outlined earlier. The ongoing was then implemented.

Add/Remove/Modify Products

The customer support department employees were trained in the process to add, remove, and modify product information as outlined earlier. The ongoing process was then implemented.

Support E-Commerce Web Site

The ENG customer service department would be responsible for fielding client calls about the usage of the new e-commerce environment, about recommendations for changes or enhancements to the environment, as well as product search assistance to clients that cannot find the products they need. A list of proposed enhancements would be created to be incorporated into the planning on the next release of the site. Any data errors identified would be resolved within one business day.

MARKETING STRATEGY FOR THE E-COMMERCE SITE

The company was faced with a tight budget as it began to plan the marketing strategy for the new e-commerce site. Both online and traditional marketing campaigns were planned and implemented to maximize the exposure of the new dental supply Web site to its target audience segments. The company embarked on a low-budget creative marketing strategy to help increase its Web site exposure. Online activities included e-mail marketing, search engine advertising, cross linking, additional online store advertising, banner advertising, online article publication in online publications and newsgroups. Traditional marketing activities included a launch event, labels on merchandise, flyers, one-on-one meetings, phone calls, letters, and advertising in paper publications.

Target Dental Market Segments

This was the first time in the company's history that it embarked on a comprehensive marketing strategy. Engaging and returning customers required a focus on the following segments:

- Existing clients of the company that were not purchasing online should be encouraged to purchase online, trained to do so, provided assistance with the initial registration process, and encouraged to purchase online on a regular basis.
- As clients started purchasing online, they were targeted to purchase more, and/or more frequently.
- Clients that used to purchase from the company in past years, but then switched to purchase from competitors and their online stores, were encouraged to return to the company by exploring its online shopping option.
- Established dental offices in the traditional geographic coverage area of the company that have been purchasing from competitors were encourages to purchase products online.
- Established dental offices outside the traditional coverage area of the company that have not purchased from the company in the past were exposed to this new online shopping option. The majority of them were not familiar with the brand, have not had a relationship or purchased in the past. This marketing exposes them to a new option for dental supply shopping.
- Newly graduated dentists from dental school are key to the future success of the organization. These were encourages to become familiar with the company, and its online store, so that they can approach it once they start working at an existing practice or once they open their own practice.
- A high-level marketing plan was created to include both online and traditional marketing methods. Parts of the plan were implemented during the first two-year period. These are discussed next.

Online Marketing Campaigns: Descriptions and Costs

Different online marketing campaigns were planned, executed, measured and evaluated by the company.

E-Mail Marketing

E-mailing is a highly effective, targeted method of reaching the dental community. An initial e-mail was sent to current to encourage them to purchase online. Regular e-mails were then sent notifying them of weekly/monthly specials. Past clients were e-mailed to show them the benefits of returning to do business with the organization via its e-commerce site. This was effective with some clients since the reason they left in the first place was the lack of ability to make online purchases.

To reach prospects via e-mail, the company first needed to determine their e-mail addresses. It is possible to extract e-mail addresses from Web sites that include discussion about dental topics, as well as dental online forums, and newsgroups. There is no cost associated with such activity beyond human resources, and this is a way to generate thousands of e-mail addresses of dentists, dental assistants, dental hygienists and others in the dental industry. An alternative is to purchase an e-mail list of all dentists in a particular region. For example, e-mails of over 8,000 dentists in southern California can be purchased for $940. E-mails are sent once per month, and recipients have the option receive weekly specials via e-mail or to unsubscribe from the monthly e-mail list, if they choose the first option, they receive weekly notification of specials via e-mail. The monthly e-mail contains a reminder to shop online, and lists a few monthly specials (those available at a discount price throughout the month). The weekly e-mail contains those products available at a discount rate for the week. Online greeting cards are e-mailed during major holidays to the subscribed e-mail list.

Search Engine Advertising

The Google AdWords program was utilized to increase traffic to the e-commerce site. The site would appear on the top right hand side of search result screens, when searching for “dental products,” “dental equipment,” “cheap dental products.” The company advertised with MSN Sponsored Sites as well.

Web Site Cross Linking

Web site cross linking was set-up with other organizations in the dental industry. Web sites of service providers such as dental labs, repair dental equipment shops, and others that cater to dentists were approached. Web sites of organizations that sell dental software, dental equipment, and provide other noncompeting products were also approached. Both parties seeking to increase their memberships benefit from such cross linking. There are no costs to this approach beyond the cost of banner design and implementation, and placement on the other organization's Web site. Possible Web sites considered for cross linking and other advertising methods include: www.handpieceheadquarters.com, www.dentalproducts.net, www.dentalxchange.com, and www.edentalmall.com.

Additional Online Store Advertising

A dental supply store was created on E-bay to expand the sale of the goods. One reason for this approach was to sell a few select popular products. More importantly, this is a way to channel more traffic from e-bay to the e-commerce Web site of the company and to enhance the company's brand. Dental supply stores were created in other related portals and online shopping malls including Amazon.com

Banners in Online Dental Publications

Research conducted as part of the company's planning efforts shows that 11% of the traditional paper publications have online versions that offer additional advertising opportunities. The Web site was linked to online portals (vertical portals of the dental industry) by paying to place banners in these Web sites that lead prospects to the company's Web site. Banner prices range from $500 per month for on the American Dental Association's Web site, to $65 per month in regional publications.

Articles in Online Dental Publications

Articles were submitted to and published in dental organizations' online newsletters, announcing the launch of the company's Web site as a way to increase exposure. Organizations included different dental association, associations of dentists, associations of dental hygienists, and associations of dental assistants.

Article Posts to Newsgroups

A search of Google Groups and Yahoo! Groups yield hundreds of groups related to dentistry. The company posted messages on these boards on a regular basis. The messages are related to the topics discussed, and suggest that additional information can be found by visiting the company's Web site. The company has also started a number of dental groups to provide people in the industry additional forums for discussion.

Traditional Marketing Campaigns in Support Dental E-Commerce

Targeted traditional marketing campaigns for the Web site were planned, executed, measured, and evaluated.

Launch Event

Local current and past clients, as well as local dentists that have not been purchasing from the company, were invited to a special PR launch event where the new e-commerce system was presented. In order to provide incentives, discounts were offered for those clients ordering products using the new system the week following the event.

Labels on Merchandise to Existing Clients

The Company's shipping department places a large label on each shipment of products to clients with the Web site address and the words "Order Online" printed on them. This approach is effectively reminds existing clients to purchase additional products online.

Flyers with Mailed Invoices

A one-page advertisement is sent with each invoice to existing clients of the dental supply company. It lists a few "Monthly Special" items that are available at discount rates only via an online purchase.

Office Visits by Sales People

Current clients of the company were visited by their sales persons to inform them of the new e-commerce capabilities and the advantages of purchasing online. Follow-up meetings were arranged, to demonstrate the Web site's functionality and create accounts for them. Clients were told that they would continue to get the levels of support expected. Sales people were asked to visit past clients, and encourage them to register online and make an initial purchase at a substantial discount. Sales people were also instructed to visit offices in their geographic coverage that are not currently serviced by the company, assist with the online registration process, and offer a substantial discount on an initial purchase.

Phone Communications with Customer Service

Customer support representatives were instructed to inform each caller about the new online shopping environment that became available. They were trained in assisting clients to go through the registration, search and purchasing processes.

Letters to Past Clients

Letters were mailed to customers that used to purchase from the company but switched to purchase from competitors in recent years.

Advertising in Paper Publications

Ads were selectively placed in a number of paper publications. A survey of the paper publications of the dental industry was conducted by the company as part of its marketing planning effort. It shows that 34% of publications are published monthly, 20% are published quarterly, 9% are annual, and the remaining 37% consist of those published bi-weekly, 3, 5, 6, 7, 9, and 10 times a year.

The average back page of a publication costs $1,336, inside back page at $882, inside front $1,112, full inside page $1,110, two thirds of a page $947, half page at $700, one third page $610, quarter page $478, one sixth of a page $318, one eighth of a page $210. The publication of the American dental association which reaches dentists through out the USA charges substantially larger advertising fees, with full page ads reading $9,030, two thirds page $6500, one half page $5,690, one third page $4,030, one fourth page $3,030, one sixth page $2,740, and one eighth page $2,450.

The average cost of advertising per publication reader is $0.56 for the back page, $0.67 for the inside back page, $0.61 for the inside front page, $0.26 for a full page ad, $0.17 for two thirds page ads, $0.16 for half page ads, $0.10 for third page as well as quarter page ads.

ANALYZING WEB SITE USAGE

Web site usage was tracked over a 2-year period as the different marketing campaigns outlined earlier were implemented.

42% of customers of the Web site are California, 45% come from other U.S. states, and the remaining come from outside of the U.S., mostly Canada and Europe.

10% of the new customers were referred to the Web site by a sales person's visit and/or phone call, 4% were referred by a colleague, 2% from attending dental shows, 12% from banner advertising, 27% from search engines, 26% from direct e-mailing, 5% from newsgroups, 10% from ads in paper publications, and 4% from other referral sources.

Most dental offices purchasing from the site are small operations of a single dentist or a couple of dentists. In small dental offices, the dentist is more likely to order dental supplies then in larger dental practices. Of the visitors to the Web site, 74% have indicated that they are dentists, while 26% indicated they are dental office staff.

During the registration process of dentists on the Web site, 51% of dental visitors have asked to receive a monthly online dental newsletter.

Of the registrants to the Web site, 34% did not login to the site during a 30 day period following registration. A call to these dentists asking them to follow-up and visit the site, resulted in 56% positive response rate. Over 19% of Web site registrants began purchasing online only following human interaction.

During the 2-year period of this study, the number of Web site user increased from zero to over 3,400 registered users. The growth was almost even, with an average of 142 registrations per month. Weaker periods of growth are exhibited during the summer months of June, July, and August, as well as the holiday season of December.

Of the shoppers, 17% did not return to purchase again on the Web site during the period, in spite of an ongoing e-mail campaign to entice them to come back. Many of these indicated price to be the main reason for their decision.

CONCLUSION AND FUTURE WORK

The decision to implement an online shopping portal for dental practitioners that serves as a gateway for accessing over 16,000 products offered by the company was important to ensure the continued existence as well as the growth of the company. The objectives of the portal were to (1) lower the cost of operations for the company, (2) improve efficiency, and (3) increase revenues.

The number of customers quadrupled and the revenues more than tripled over this 2 year period with a very limited budget.

The ability to capture demographic information, shopping habits and other vital information was also implemented. Several kinds of reports were created through data mining to perform effective targeted marketing. These will be utilized over the next few years to take the company's marketing for existing and past clients to a new level.

We believe that a human touch must accompany each client to complement and enhance their online experiences. This encourages customers to purchase more, and more frequently.

Larger dental supply companies offer more services and more products to their clients, which is what sets them a part. The company is planning to transform part of its Web site into a dental services portal, linking its visitors with repair services for hand pieces, dental office design specialists, dental office software automation vendors that will offer their clients discounts. The portal will include: (1) Office design assistance: design the office from the equipment they need to work to the furniture, computer software programs they need to operate their facility. (2) Business rewards cards: Offer business cards to customers where on each purchase with the company, they can receive a certain percentage of money back or points towards other company products. (3) Demographic site analysis: Give business advice to new dentists. The company will provide business analysis for new dentists on the local economy, future growth potential and real estate potential. (4) List of conventions and seminars: The company will be prepared with a list of upcoming conventions and seminars for new dentists, veteran dentists, dental hygienist, dental assistants and dental students. (5) Jobs portal to attract those in the dental profession looking for a new job. The portal would provide the company with additional reasons to interact with prospective buyers, inform them of the company and its products, and convert them to shoppers of its online shopping environment.

REFERENCES

Bauer, J.C., & Brown, W.T. (2001). The digital transformation of oral health care - teledentistry and electronic commerce. *Journal of the American Dental Association*, *132*(2), 204-209.

Benco. Retrieved February 15, 2008, from http://www.benco.com

Goff, S. (2002). Internet purchasing's growing popularity. *Dental Lab Products: Buyers Guide 2002*, *27*(2), 24.

Hirschinger, R. (2001, March). Digital dentistry: Information technology for today's (and tomorrow's) dental practice. *Journal of the California Dental Association*, *29*(3), 215-221; 223-225.

Marks, R.G. (2004). The future of Web-based clinical research in dentistry. *Journal of Dental Research, 83*, C25-28C.

Palmer, D. (2002). Organizational adoption of a new e-commerce innovation: A controlled field experiment within the market channel. *Proquest Information and Learning Company,* 1-22.

Patterson Dental Supply. Patterson Companies Investor Relations. Retrieved February 15, 2008, from http://equitymarketpartners.com/PDCO

Pearson. Retrieved February 15, 2008, from http://www.pearson-dental.com

Schleyer, T.K., & Spallek, H. (2002). An evaluation of five dental Internet portals. *Journal of the American Dental Association, 133*(2), 204-212.

Schleyer, T.K., Spallek, H., Bartling, W.C., & Corby, P. (2003, January). The technologically well-equipped dental office. *Journal of the American Dental Association, 34*(1), 30-41.

Schleyer, T.K., Spallek, H., & Torres-Urquidy, M.H. (1998, December). A profile of current Internet users in dentistry. *Journal of the American Dental Association, 129*(12), 1748-1753.

Sullivan-Schein. Retrieved February 15, 2008, from http://www.henryschein.com

U.S. Department of Commerce. (2000). *The U.S industry and trade outlook - medical and dental instruments and supplies.* The McGraw-Hill Company (pp. 1-10).

X-Cart Software Solutions. Retrieved February 15, 2008, from http://www.x-cart.com/

Creating Patient-Centered E-Health

E. Vance Wilson
University of Toledo, USA

INTRODUCTION

As e-health applications have increased in number and variety, the generalized concept of e-health as "health services and information delivered or enhanced through the Internet" (Eysenbach, 2001) has lost much of its value as a mechanism for guiding development and research in this emerging field (Pagliari et al., 2005). E-health has expanded to comprise purely clinical applications (e.g., physicians consulting on a diagnosis) (Wiecha & Pollard, 2004), emergency health communication applications (e.g., for distributing information about SARS) (Rizo, Lupea, Baybourdy, Anderson, Closson & Jadad, 2005), disease-focused applications (e.g., diabetes self-management support) (McKay, Glasgow, Feil, Boles & Barrera, 2002), applications to support electronic communication between patients and physicians (Wilson, 2003), and commercial applications that have no association with a patient's own health care provider (e.g., WebMD) (Itagaki, Berlin & Schatz, 2002). It is clear that the needs of both users and researchers vary widely across these diverse applications, and I argue that both groups would benefit from development of finer-grained approaches to thinking about e-health.

BACKGROUND

Initial e-health applications were designed from a *consumer-centered* perspective, drawing upon an e-commerce business model in which vendors expected to profit from consumers paying directly for products and services they would acquire through the e-health Web site or from advertisers paying for exposure to consumers (Wilson, 2006). Most of these commercial e-health ventures failed during the dot-com bust (Itagaki et al., 2002). However, health care provider organizations have emerged to fill the void by supplying *provider e-health* to their own patients,[1] including generalized informational content as well as appointment scheduling, prescription refilling, and online communication with physicians and clinical staff (Lazarus, 2001; Wilson & Lankton, 2004).

Despite the goal of serving patients, most provider e-health has been developed from an *organization-centered* perspective that is based upon the provider's needs and economic interests. Patients' needs tend to be based on assumptions made by e-health developers and provider administration rather than being driven by patients themselves (Wilson, 2006). As a result, patients sometimes ignore e-health applications that providers assume will be the most essential to their interests. For example, Payton and Brennan (1999) describe e-health that provided Alzheimer caregivers a sophisticated, interactive, decision-support utility, e-mail communication with a nurse-moderator, an electronic bulletin board, and an online medical encyclopedia. The decision support system was considered by developers to be essential to caregivers and was substantially more expensive to deliver than other components. Yet monitoring over an 18-month period showed that decision support was used less than one-tenth as frequently as simple online communication services consisting of an electronic bulletin board and e-mail communication with a nurse-moderator.

CENTERING ON THE PATIENT

This chapter approaches e-health from a *patient-centered* perspective. Patients comprise a large and growing constituency of e-health users (Krane, 2005), and surveys indicate there is very high interest among patients in increased access to provider e-health for a variety of specific interaction needs (e.g., physician-patient communication) (Taylor & Leitman, 2002). Leaders in the medical community are coming to recognize that patients *expect* to be empowered in making health care decisions (Institute of Medicine, 2001), and that the expectation of personal control is especially strong for e-health applications (Lafky, Tulu & Horan, 2006; Markle Foundation, 2004). However, health care provider organizations have only recently begun to provide patients with online access to health care services, and numerous obstacles are present that can block development of effective patient-centered e-health applications. These include:

1. Financial disincentives for participation by outside parties (i.e., participants in e-health other than the patient, including physicians and provider administration)
2. Reduced work quality resulting from participation by outside parties
3. Reluctance of outside parties to relinquish control to patients
4. Restrictive interpretation of privacy and security regulations
5. Discomfort of patients and other key parties with computing environments.

A complete discussion of means for overcoming such obstructions is well beyond the scope of this chapter (for such a discussion, see Tan, Cheng & Rogers, 2002). Potentially, obstructions can be overcome as has been the case for other heavily regulated, traditional industries (e.g., banking and financial services), which have successfully transitioned to the online environment.

Guiding Principles for Creating Patient-Centered E-Health

In order to meet patients' expectations, it is essential for developers to focus on several guiding principles that are distinct from alternative approaches. Specifically, these are:

1. Focus on desired interactions in which the patient is an active participant
2. Incorporate only those services that meet the expressed needs of patients or are validated against patient needs
3. Be understandable to patients
4. Provide easy access for patients to completely manage and control functionality
5. Provide ready interoperability to support interaction with outside parties (e.g., physicians and pharmacy) and with other health information systems (e.g., hospital billing)

These principles correspond to a large extent with user-centered design principles that have proved important to the success of developing Web applications across numerous contexts outside the health care domain (Lazar, 2001).

The first principle emphasizes patient *involvement*, thereby distinguishing patient-centered e-health from other applications such as telemedicine, where the patient is primarily an object of the interaction rather than an active participant. The second principle addresses patient *interest* and specifically cautions against relying on untested assumptions about patients as a basis for e-health design. By emphasizing patients' involvement and interest, the first two principles help to ensure that patients will have inherent motivation to use related e-health applications. They also guide the process of eliciting patients' interaction needs and mapping these to e-health services. The remaining principles center on accessibility and source of control.

Principle 3 proposes that e-health information and communication should be understandable to patients. Some researchers argue that it is patients' own health literacy that must increase in order for e-health to succeed (Norman & Skinner, 2006). However, this type of idealistic mindset ignores first that patients' *need* for health care services is not dependent upon their literacy level, and second, that patients can benefit greatly if e-health designers make the effort to incorporate simple explanations and illustrations where these are practical. Many individuals who are only marginally literate have proved to be highly capable of interacting with online applications (e.g., banking) when they are provided with effective technological support. Further, requiring patients to be highly literate in order to use e-health is no more defensible logically than requiring high literacy in order to schedule exams or other health care services that the provider may offer. From the patient's perspective, e-health is simply an extension of the providers' other services; thus, it is reasonable for patients to expect e-health to be generally understandable and for the provider to offer mechanisms by which better explanations can be obtained if these are needed.

Principle four presents a clear statement that ultimate control of patient-centered e-health must flow to the patient. Increasing patients' involvement in e-health has been promoted recently as part of the U.S. plan for "delivering consumer-centric and information-rich health care" (Thompson & Brailler, 2004). However, achieving patient control faces two key obstacles. First, medical institutions are only slowly moving away from a firmly embedded authoritarian or "paternalistic" model of physician-patient relationships (Emanuel & Emanuel, 1992) in which physicians expect to control virtually all aspects of their interaction with patients (Eysenbach & Jadad, 2001). Embedded cultural practices can take significant time to change, even in the face of substantial social pressures. Therefore, it is likely in the foreseeable future that many physicians and other outside parties will resist allowing patients to exercise complete control over e-health or other aspects of health care. Second, health care provider organizations are reluctant to open up health information systems (HIS) to access by patients. Reluctance is based on several factors, none of which is necessarily unreasonable. Providers have the responsibility to maintain privacy and security of patient and provider data, which could be compromised by increasing accesses to HIS. Significant labor expenditures will likely be necessary in order to interconnect HIS with e-health, and it may be difficult for providers to identify ways that these expenses can be offset by increased income. Furthermore, access to HIS may be structurally blocked by disparate storage and communication formats among various proprietary systems. Lack of access to the provider's HIS or to specific parts, such as medical records and test results, leaves patients with relatively little to control, as the e-health application will not be able to offer direct interaction with key functions, such as billing and appointment scheduling. Therefore, it is important to develop strategies that can economically enhance access to HIS without compromising security or can establish alternatives to direct access (e.g., applying a data warehouse model where only duplicate data are accessible to patients). It also is key to promote among physicians and clinical staff that patients have the right to play a central role in their own health care decisions.

Principle five emphasizes that interaction with outside parties is essential for creating effective patient-centered e-health. Technical approaches can be important to achieve interoperability, as demonstrated by Wilson and Lankton (2003) in their proposal of a guided-mail system as an alternative to e-mail for patient-physician communication. However, social issues should not be overlooked in achieving interoperability, as these may be even more important than technical aspects in motivating outside parties to participate. For example, patients embrace the idea of having access to their own medical records, yet physicians are skeptical that patients would benefit from access and worry that their workload would increase if records became available (Ross, Todd, Moore, Beaty, Wittevrongel & Lin, 2005).

By addressing these five principles, developers can work to create e-health that is centered on patients' needs as well as accessible and meaningful in its capabilities. It may be anticipated that other design principles will also apply (e.g., requirements arising from privacy and security regulations). However, care should be taken to ensure that these do not unnecessarily obstruct any of the five principles of patient-centered e-health that are enumerated previously.

Analyzing Patients' Interaction Needs

This section demonstrates a method for analyzing patient interaction needs. Recall that the principles—*(1) focus on desired interactions in which the patient is an active participant* and *(2) incorporate only those services that meet the expressed needs of patients or are validated against patient needs*—are intended to guide the process of eliciting patients' desired interactions and mapping these to e-health services. Both principles require e-health developers to interview representative patients. Interviews may be conducted with individual patients or in facilitated focus groups. In both settings, the key objectives are (1) to establish what participants the patient desires to interact with regarding his or her own health care and (2) to list the specific types of interaction the patient needs to perform with each participant. A third objective is to discuss specific e-health services that could meet each interaction need and assess the patients' evaluations of these. This discussion typically is conducted at the end of the interview or focus group session.

The sample patient interaction chart shown in Table 1 illustrates results of a completed interview in which each of the objectives is represented by a column. Column 1 identifies the desired participants (other than the patient) involved in health care interactions. Column 2 describes the interaction need, and column 3 maps that interaction need onto an e-health service. Note that in several cases interactions map to the same type of e-health service. For example, most interactions that access generic information can be satisfied by linking to appropriate static Web pages (i.e., Web links). Other services may be shared, assuming that e-health interconnects with the specified other participant. Making

Table 1. Sample patient interaction chart

Other Participant	Interaction Need	E-Health Service
Self	Manage health information	Personal health record
	Set up and receive personal health reminders	Reminder system
Other patients	Participate in peer support groups	Discussion lists
Physicians and clinical staff	Access electronic medical record (EMR)	EMR interface
	Communicate with physicians and staff	Online mail
	Receive reminders and office visits follow-up	E-health mail
	Connect for remote diagnosis and monitoring	Instrumentation interface
	Order prescription refills	Prescription manager
	Request referrals	Referral request
Provider administration	Schedule appointments	Appointment scheduler
	Manage co-pays and financial accounts	Payment center
	Communicate with administrative staff	E-health mail
	View provider's promotional and educational materials	Web links
Pharmacies	Manage prescription ordering, delivery, and pick-up	Prescription manager
	Manage co-pays and financial accounts	
	Research medication instructions, side-effects, and	Payment center
	potential interactions	Online drug formulary
Supporting facilities (lab, rehabilitation center, in-home services, etc.)	View patient instructions and facility policies	Web links
	Schedule appointments	Appointment scheduler
	View test results	Web links
Employer health services	Manage health plan enrollment forms	Web links
	Research alternative insurers	Web links
	View employer's promotional and educational health materials	Web links
	Receive notifications of changes to health plan options	E-health mail
Insurers	Manage insurance forms	Insurance manager
	Manage financial accounts	Payment center
	Solicit pre-approvals for procedures and medications	Insurance Manager
	Research participating physicians	
		Web links
Governmental agencies	Manage school health forms	Web links
	Research government services	Web links
	Receive public health notifications and alerts	E-health mail

payments or sending/receiving health-related e-mail can be handled efficiently through a shared clearinghouse service, regardless of which other participants these involve.

It is important for developers to recognize that interaction participants and needs can vary widely among patients, depending upon socioeconomic class, health condition, and age. Patient needs should dictate how participants are categorized and what level of e-health services are provided. For example, geriatric patients who are dependent upon government programs may need detailed interactions with governmental agencies. In this case, the governmental agencies participant category in Table 1 may need to be expanded to represent several agencies.

Personal Health Informatics and Patient-Centered E-Health

Creating effective patient-centered e-health is an important goal, but as discussed previously, this goal faces numerous obstacles, including fears of financial disincentives and reduced work quality for outside parties who are called upon to participate. Key to these concerns is the current lack of a focused discipline in which creation of patient-centered e-health is a core responsibility. It will remain difficult to champion patient-centered e-health initiatives until such a discipline coalesces through establishment of training programs, identification of best practices, and development of a shared identity.

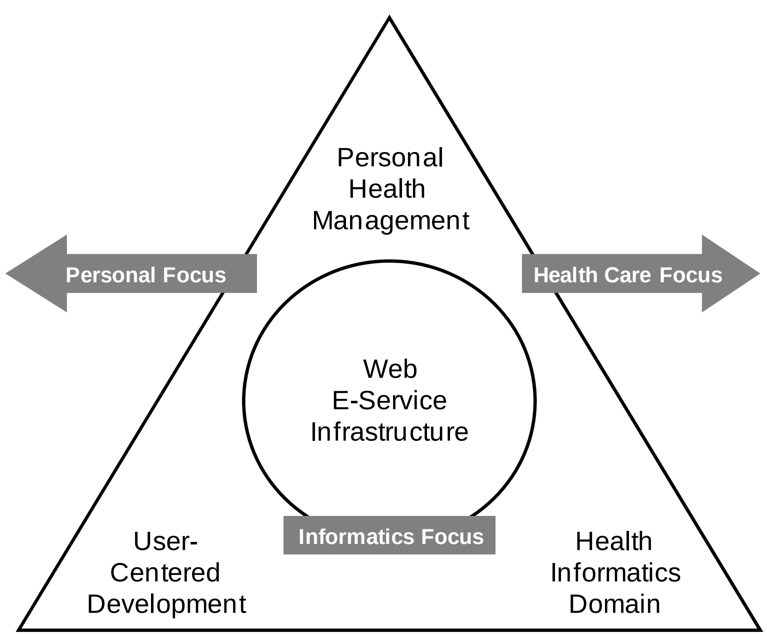
One training approach that satisfies the major goals and guiding principles of patient-centered e-health is the Personal Health Informatics (PHI) agenda proposed by Wilson (2006). PHI incorporates the knowledge, skills, practices, and research perspectives necessary to develop patient-centered e-health. As shown in Figure 1, four essential content areas support PHI, and these are applied within three focal areas. *Web e-service infrastructure* describes the hardware, software, and networking capabilities that support e-health functions. In recognition that e-health is primarily service-oriented, the *informatics focus* in PHI centers on infrastructure that is specialized for e-service presentation and delivery. *Personal health management* and *user-centered development* make up the *personal focus* that PHI presents to patients. User-centered development methods provide tools for eliciting user needs, designing solutions, and evaluating the utility of these solutions in meeting patients' needs. Personal health management addresses patients' individual practices as well as psychological, social, and cultural aspects of their management of personal health information. Personal health management and the *health informatics domain* combine in the *health care focus* of PHI. Content drawn from the health informatics domain addresses the skills and knowledge necessary to interface with HIS and related health care systems.

Development of a PHI discipline in which personnel train in these four essential content areas and share a joint focus on patients (personal), health care, and informatics would provide substantial leverage toward achieving effective patient-centered e-health. Patient surveys consistently show a groundswell of desire for e-health applications that meet patients' specific needs, including needs for electronic support for managing personal health information (Taylor, 2004), patient-physician communication (Taylor & Leitman, 2002), and high-quality online health information (Krane, 2005). Approaches such as PHI will be essential in concentrating the skills, knowledge, and patient-centered perspective that are necessary to meet these growing patient demands.

CONCLUSION

The emergence of e-health has been overwhelmingly beneficial to patients, who already conduct many types of transactions and information searches on the Internet and strongly wish to be able to access health care services and information in the same manner. I argued in this chapter that patients comprise a large and growing constituency of e-health users whose interests would be served by creating e-health from the patient's perspective. The guiding principles and interaction need

Figure 1. Conceptual model of key PHI structural components (Wilson, 2006)



analysis techniques I have outlined important steps toward achieving patient-centered e-health. However, there remains a need to develop an academic and professional discipline in which patient-centered e-health constitutes a core responsibility, as represented by the PHI agenda presented here.

REFERENCES

Emanuel, E.J., & Emanuel, L.L. (1992). Four models of the physician-patient relationship. *JAMA, 267*(16), 2221–2226.

Eysenbach, G. (2001). What is e-health? *Journal of Medical Internet Research, 3*(2), e20.

Eysenbach, G., & Jadad, A.R. (2001). Evidence-based patient choice and consumer health informatics in the Internet age. *Journal of Medical Internet Research, 3*(2): e19.

Institute of Medicine (2001). *Crossing the quality chasm.* Washington, DC: Institute of Medicine, National Academies Press.

Itagaki, M.W., Berlin, R.B., & Schatz, B.R. (2002). The rise and fall of e-health: Lessons from the first generation of Internet healthcare. *Medscape General Medicine, 4*(2). Retrieved August 10, 2006, from http://www.medscape.com/viewarticle/431144_Print

Krane, D. (2005). Number of "cyberchondriacs"—U.S. adults who go online for health information—increases to estimated 117 million. *Healthcare News, 8*(5). Retrieved December 13, 2005, from http://www.harrisinteractive.com/news/newsletters_healthcare.asp

Lafky, D.B., Tulu, B., & Horan, T.A. (2006). A user-driven approach to personal health records. *Communications of the Association for Information Systems, 17,* 1028–1041.

Lazar, J. (2001). *User-centered Web development.* Sudbury, MA: Jones and Bartlett.

Lazarus, I.R. (2001). Separating myth from reality in e-health initiatives. *Managed Healthcare Executive,* June, 33–36.

Markle Foundation. (2004). *Achieving electronic connectivity in healthcare.* Markle Foundation.

McKay, H.G., Glasgow, R.E., Feil, E.G., Boles, S.M., & Barrera, M. (2002). Internet-based diabetes self-management and support: Initial outcomes from the Diabetes Network Project. *Rehabilitation Psychology, 47*(1), 31–48.

Norman, C.D., & Skinner, H.A. (2006). eHealth literacy: Essential skills for consumer health in a networked world. *Journal of Medical Internet Research, 8*(2), e9.

Pagliari, C., et al. (2005). What is eHealth (4): A scoping exercise to map the field. *Journal of Medical Internet Research, 7*(1), e9.

Payton, F., & Brennan, P. (1999). How a community health information network is really used. *Communications of the ACM, 42*(12), 85–89.

Rizo, C.A., Lupea, D., Baybourdy, H., Anderson, M., Closson, T., & Jadad, A.R. (2005). What Internet services would patients like from hospitals during an epidemic? Lessons from the SARS outbreak in Toronto. *Journal of Medical Internet Research, 7*(4), e46.

Ross, S.E., Todd, J., Moore, L.A., Beaty, B.L., Wittevrongel, L., & Lin, C. (2005). Expectations of patients and physicians regarding patient-accessible medical records. *Journal of Medical Internet Research, 7*(2): e13.

Tan, J., Cheng, W., & Rogers, W.J. (2002). From telemedicine to e-health: Uncovering new frontiers of biomedical research, clinical applications & public health services delivery. *Journal of Computer Information Systems, 42*(5), 7–18.

Taylor, H. (2004). Two in five adults keep personal or family health records and almost everybody thinks this is a good idea: Electronic health records likely to grow rapidly. *Health Care News, 4*(10). Retrieved December 13, 2005, from http://www.harrisinteractive.com/news/newsletters_healthcare.asp

Taylor, H., & Leitman, R. (2002). Patient/physician online communication: Many patients want it, would pay for it, and it would influence their choice of doctors and health plans. *Health Care News, 2*(8). Retrieved December 13, 2005, from http://www.harrisinteractive.com/news/newsletters_healthcare.asp

Thompson, T.G., & Brailler, D.J. (2004). *The decade of health information technology: Delivering consumer-*

centric and information-rich health care. Washington, DC: Department of Health and Human Services.

Wiecha, J., & Pollard, T. (2004). The interdisciplinary eHealth team: Chronic care for the future. *Journal of Internet Medical Research, 6*(3), e22.

Wilson, E.V. (2003). Asynchronous health care communication. *Communications of the ACM, 46*(6), 79–84.

Wilson, E.V. (2006). Building better e-health through a personal health informatics pedagogy. *International Journal of Healthcare Information Systems and Informatics, 1*(3), 69–76.

Wilson, E.V., & Lankton, N.K. (2003). Strategic implications of asynchronous healthcare communication. *International Journal of Healthcare Technology and Management, 5*(3/4/5), 213–231.

Wilson, E.V., & Lankton, N.K. (2004). Modeling patients' acceptance of provider-delivered e-health. *Journal of the American Medical Informatics Association, 11*(4), 241–248.

KEY TERMS

Consumer-Centered e-Health: E-health applications created to generate profit through sales of products, services, and/or advertising.

e-Health: Health services and information delivered or enhanced through the Internet (Eysenbach, 2001).

Organization-Centered e-Health: E-health applications created to meet organizational needs and economic interests.

Patient-Centered e-Health: E-health applications created to support interactions that are desired and needed by patients.

Patient Interaction Needs: Interactions with other individuals or institutions that the patient desires and considers to be needed.

Personal Health Informatics: The knowledge, skills, practices, and research perspectives necessary to develop patient-centered e-health.

Provider e-Health: E-health applications supplied by a health care provider organization primarily for use by its own patients.

C

Critical Issues in Assessing Sustainability and Feasibility of E-Health

Nilmini Wickramasinghe
Illinois Institute of Technology, USA

Santosh K. Misra
Cleveland State University, USA

INTRODUCTION

The cost of health care is increasing exponentially worldwide. The adoption and diffusion of e-health and the application of Internet and Communication Technology (ICT) in health care is growing at a rapid rate in an attempt to find cost-effective methods of providing quality health care. Both European and US governments are making e-health a priority on their agendas. However, few, if any, discuss the critical issues of the sustainability and feasibility of e-health models. We attempt to fill this critical void by presenting a macro framework that identifies the key components of a generic e-health system and identifying factors playing a role in the assessment of e-health sustainability.

BACKGROUND

During his State of the Union Address in January 2004, President George W. Bush affirmed the intention of the government to emphasize the role of technology in administration and delivery of health care in the United States (Bush, 2004). Similar sentiments have been voiced by European leaders (Global Medical Forum Foundation, 2005; Oslo Declaration on Health, 2003) and the World Health Organization (WHO) (e-health in eastern Mediterranean, 2005; A Health Telematics Policy, 1998). Both European and US authorities define their initiatives primarily in terms of medical information technology, centering on computerized patient records (CPRs) or, in more acceptable parlance, the EHR (electronic health record) (Brailer & Terasawa, 2003). WHO's platform statement (A Health Telematics Policy, 1998) speaks of a "health telematics policy," an all-inclusive term that incorporates not only EHR but also all health care services provided at a distance and based on the use of information and computer and communications technologies (IC^2T). Countries are now turning to various e-health solutions in order to stem the escalating costs of health care and yet provide superior health care delivery.

As any student of electronic business (e-business) knows, a number of dot.com companies went out of business in the early 2000s. Going out of business is also a fairly regular phenomenon in the brick and mortar world. Health care organizations have not escaped this "going out of business" problem. It is expected that e-health will face similar going out of business issues. At the same time, many businesses, including many health care organizations, have maintained continuity of profitable operations, and it stands to reason that e-health organizations can also do the same. Sustainability of businesses has been studied extensively by management theorists. However, to date, few studies, if any, have attempted to assess the sustainability and feasibility of e-health.

In this chapter, we attempt to study the sustainability of e-health. For this, we draw from our work on assessing the sustainability and feasibility of e-business models in general (Sharma, Wickramasinghe & Misra, 2002). Our proposed sustainability macro model also builds on our previous work in identifying the competitive forces facing e-health (Wickramasinghe, Misra, Vogel & Jenkins, 2006).

IDENTIFYING KEY ISSUES OF SUSTAINABILITY IN E-HEALTH

E-Business Models

The prime objective of a business model is to ensure that the organization is profitable (La Monica, 2000). E-businesses that offer products and services online must work efficiently and effectively to create profit margins. E-businesses, like any business, can compete in the marketplace by differentiating their products and services from their competitors and by providing unique and better customer value. A product or service can be differentiated if customers perceive some value in the firm's products and services that competitors cannot match. Differentiation also can be done by offering different product features, timing, location, service, product mix, and linkages among functions (Afuah & Tucci, 2000). One can judge the customer value by examining whether a firm is offering its customers something distinctive at a cost the customer is willing to pay, and whether the combination of distinctiveness and price is a better value than provided by the firm's competitors. The success of the business model would then depend upon how the firm sets a price for the value offered.

Pricing and profitability are factors of market shares and margins. A good business model can strive for high market share and devise strategies to maintain the market share. The cost and asset model of a firm also influences the product pricing decisions. The cost (fixed cost + variable cost) should be spread out in an appropriate fashion in order to maintain positive profit margins. For electronic businesses, profit margin is not only derived from product sales but also from many other revenue sources. Such revenue sources may include banner advertising, membership, and referral fees. In addition, a firm can have significant competitive advantage if it offers products and services that cannot be imitated easily. The sustainability of a firm can be judged by examining issues associated with maintaining and improving the firm's market share as well as sources of its competitive strength. For example, using a simple profit equation (Profits=(P-Vc)Q-Fc+OthRev), we can examine how each of the components of a business model impacts profitability. If a firm offers distinctive products, it can charge a premium price P and increase the margin. The firm also can increase profit by reducing the variable cost Vc by increasing the market share Q and other revenues (OthRev) such as banner advertising fees (Afuah & Tucci, 2000).

E-Health Business Model

Our previous work (Figure 1) (Wickramasinghe et al., 2006) (1) has served to map the key components of an e-business model into the health care domain. Specifically, this model incorporates all the key actors of health care, including suppliers, health care organizations, providers, regulators, payers, and, of course, the patient. It is important to note that the dynamics of health care are quite unique (Gargeya & Sorrell 2004; Wallace, 1997; Wickramasinghe et al.,

Figure 1. E-health business model components

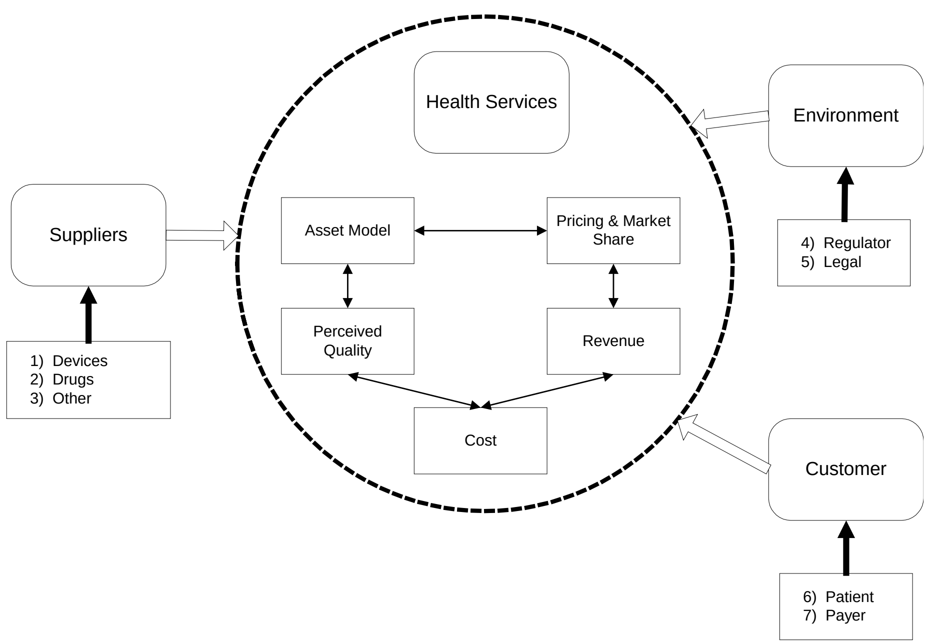


Table 1. Key areas of sustainability

Competitive Positioning	Product/Service Competition and Industry Structure Health care Consumer Value
Economic Positioning	Revenue Model Cost Model
Enterprise Resource	Asset Model People and Partners Processes Technology

2006) and that health care delivery by its very nature necessitates people-to-people interactions. E-health not only requires a physical organization (the brick-and-mortar component) but also a supporting e-health network. This idea of a netcentric perspective to health care already has been established by von Lubitz and Wickramasinghe (2006). Given the netcentric nature of e-health, sustainability of e-health is essentially the sustainability of the e-health network. We consider this distinction to be especially important since there is a near universal move to embrace various e-health initiatives in an attempt to offer superior, cost-effective health care delivery.

Key Elements of Sustainability

In an elementary sense, an enterprise is sustainable when it can maintain a stream of profit over a long period of time. As shown in the case of e-business, sources of profit are many and can be summarized into three primary areas: competitive positioning, economic positioning, and enterprise resources of a business (Afuah & Tucci, 2000; Wickramasinghe et al., 2006). Table 1 identifies a few of the important factors that influence each of these key areas. We argue that any sustainability model must take all these areas and their contributing factors into consideration (Sharma et al., 2002).

In addition, it is also prudent to consider health care service demand as a key area. If the demand for health care services through e-health does not exist or is very small, then an e-health network would find it much more difficult to sustain itself.

In Table 2, we examine in more details some of the primary components of the key areas: demand, revenue, cost, and asset.

E-Health Sustainability

Figure 2 proposes a macromodel for an e-health network's sustainability. The model components shown in Table 2 are mapped into the model of Figure 2. The macromodel of Figure 2 also shows interdependence among model components.

An e-health network can be sustained if it can maintain positive revenue margin. Other key measures of sustainability include market share, retained earnings, and growth rate. A positive margin is generated when revenues exceed cost. As we have shown in Table 2, there are many potential sources of revenue as well as many different cost drivers. The demand for e-health services is potentially global and has a significant impact on the revenue stream. The asset model as well as the cost model of e-health impact the products and services of the e-health network. Finally, the asset, cost, and revenue models themselves benefit from the positive revenue and market shares.

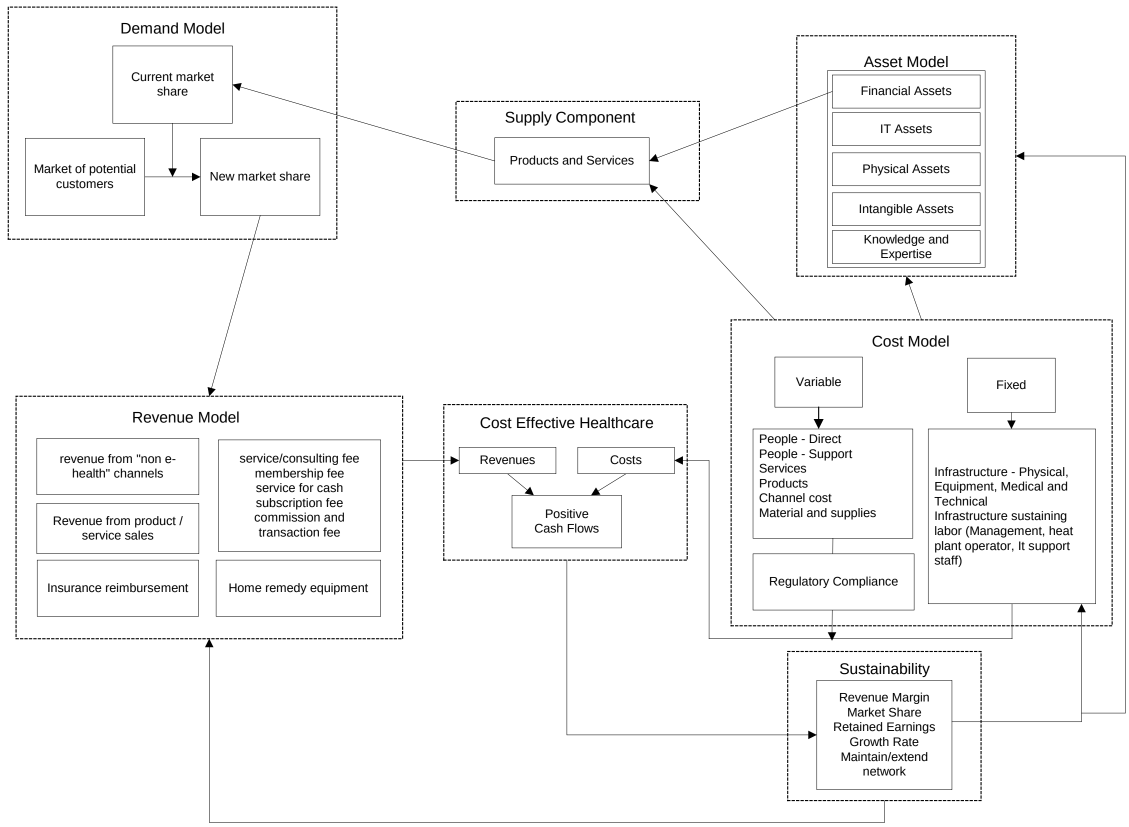
FUTURE TRENDS

Our e-health sustainability model in Figure 2 has many significant implications to health care. First, it is important to note that e-health sustainability links to sustainability of the country's health care model. This means that any e-health initiative must be considered within the context of the health care polices and structures of that country. The e-health sustainability model underscores the importance of e-health as providing a network for which superior cost-effective health care delivery can ensue. Moreover, the e-health sustainability model represents a macro sustainability model, and it is critical that in the development of any e-health initiative, a holistic systems perspective is adopted. Within this

Table 2. Key model components for assessing sustainability (adapted from Sharma et al., 2002)

Model Component	Description
Demand	
Demand Type	
Current Market Share	The market share enjoyed in a pre- e-health network state. An existing organization would have a market presence that defines the baseline from which growth can occur due to e-health.
Market of Potential Customers	The larger market that can be realized
New Market Share	The market share that results from entry into e-health segment of health care. This growth should occur since a health care consumer would be able to benefit from a complementary channel of service delivery. For those organizations with a brick-and-mortar presence, e-health is an additional channel; for purely e-health networks, any gain in market share is new market share.
Revenue	
Revenue Type	
Product or Services Sales	The revenue from the sales of goods or services.
Advertising	The fee for hosting other companies' advertisements
Subscription Fee	The fee from subscriptions to receive certain products or services
Membership Fee	The fee for memberships in interest groups
Commissions or Transaction Fee	The transaction fee or commission as an agent or broker for facilitating the selling of partner organizations' products and services
Insurance Fee	Reimbursement from insurance providers for services rendered
Cost	
Cost Type	
People	Salaries and other benefits of employees and consultants. This includes cost of hiring, developing, and training as well as operational costs.
Advertising or Marketing Expenses	Cost for identifying health care consumers and markets. This cost also includes all channel costs to deliver nonelectronic services.
Commissions or Transaction Fees Cost	Paying commissions or transaction fees to agents or brokers who are selling a company's products and services
Material and Supplies	Cost to procure materials and supplies
Physical infrastructure	Cost for facilities, equipment, buildings, utilities, etc.
IT Infrastructure	Cost of computer setup, network costs, Internet subscription, development costs, and operating and maintenances costs
Asset Model	
Asset Type	
Financial Assets	Balance sheet, cash flows
Physical Assets	Plants, equipment, etc.
IT Assets	Information technology infrastructure
Intangible Assets	Relationships with partner organization, agility and strength of brand, etc.
Intellectual Assets	Knowledge and expertise of key participants, patents and copyrights, tacit knowledge.

Figure 2. Macromodel of sustainability



network, each player who contributes to the network supplier, health care organization, provider, or payer must then look at its own organization's micro sustainability model, while the patient must make judicious and informed choices for the specific care treatment; the role of the regulator should be that effective governance of this system is maintained at all times.

The model presented in Figure 2 is a generic business model applicable to many e-health business scenarios. At this point, we have not made any distinction in terms of an e-health organization that operates as a stand-alone virtual organization or as an additional channel to an existing health care entity. While we argue that the sustaining factors would be present in both of these groups of e-health, we believe the intensity of contribution would be significantly different based on organizational structure for a specific e-health entity. In addition, we hypothesize that a trust and reputation relationship would play a significant role in the long-term sustainability of an e-health organization, even though this role is implied within the intangible asset and knowledge/expertise components of the asset model defined in Figure 2. In our opinion, these issues require further exploration.

CONCLUSION

Business sustainability is not a new concept. It has been studied by management theorists for a long time. The structure of an e-health network has many unique features not ordinarily associated with a typical business. Nevertheless, an atypical business such as e-health must still be sustainable. In this chapter, we analyze the factors that play a role in sustaining this atypical organization. A macromodel is proposed. Components of this model are identified, and their interdependence is mapped. Such a model is important as countries and health systems try to incorporate various e-health initiatives and also as health care infrastructures are rebuilt after devastation; for example, the tsunami affected countries of the Indian Ocean, or war affected countries such as Iraq.

We propose to refine this model in our future studies. We also propose to derive empirical relationships among model components so that predictive measures for sustainability of e-health can be developed.

REFERENCES

Afuah, A., & Tucci, C.L. (2000). *Internet business models and strategies—Ttext and cases*. McGraw-Hill Irwin.

Brailer, D.J., & Terasawa, A.B. (2003). Use and adoption of computer-based patient records. California HealthCare Foundation, 1–42.

Bush, G.W. (2004). State of the union address. http://www.whitehouse.gov/news/release/2004/01/20040120-7.html

e-Health in the eastern Mediterranean, WHO-EMRO (2005). http://www.emro.who.int/his/ehealth/ehealthPlan.htm

Gargeya, V., & Sorrell, D. (2004). Moving toward an e-hospital. In N. Wickramasinghe, J. Gupta, & S. Sharma (Eds.). *Creating knowledge-based healthcare organizations* (pp. 50–64). Hershey, PA: Idea Group Publishers.

Global Medical Forum Foundation. (2005). GMF IV: New frontiers in healthcare. Retrieved April 4–6, 2005, from http://www.globalmedicalforum.org/gmf_iv/session3.asp

World Health Organization. (1998). A health telematics policy in support of WHO's health-for-all strategy for global health development: Report of the WHO group consultation on health telematics. Geneva: World Health Organization.

Hofer, C. (1975). Toward a contingency theory of business strategy. *Academy of Management*, *18*(4), 784–810.

La Monica, M. (2000). Building trust into e-business models. *InfoWorld, 22*(28):3.

The Oslo Declaration on Health, Dignity and Human Rights. (2003). Proceedings of the 7th Conference of European Health Ministers, Oslo, Norway. Retrieved from http://odin.dep.no/hod/norsk/aktuelt/taler/monister/042071-990209/dpk-bn.html

Sharma, S., Wickramasinghe, N., & Misra, S. (2002). *A framework to assess the feasibility and sustainability of electronic business models*. Proceedings of the 5th ICECR Conference, Montreal.

Thorne, K., & Smith, M. (2000). Competitive advantage in world class organizations. *Management Accounting*, *78*(3), 22–26.

von Lubitz, D., & Wickramasinghe, N. (2006). Healthcare and technology: The doctrine of networkcentric healthcare [in press]. *Intl. J Electronic Healthcare (IJEH)*.

Wallace, S. (1997). Health information in the new millennium and beyond: The role of computers and the Internet. *Health Education*, *3*, 88–95.

Wickramasinghe, N., Misra, S., Vogel, D., & Jenkins, A. (2006). The competitive forces facing e-health [forthcoming]. *Intl J Healthcare Information Systems and Informatics.*

KEY TERMS

Asset Model: Components of business model that define the assets owned by the business entity.

Cost Model: Components of business model that define the cost sources for the organization.

Demand Model: Components of business model that define the customer needs and wants.

E-Health: The offering of health care products and services using Web-based technologies.

E-Business Model: A business model where either the product or process or selling agent is partially or completely electronic.

Intangible Asset: Assets such as goodwill, reputation, and brand recognition.

Revenue Model: Components of business model that define the sources of revenue for the organization.

Sustainability: The long-term survivability of the business entity.

C

Critical Role of Bionanotechnology in Healthcare and Society

Shivani Choudhary
Illinois Institute of Technology, USA

Nilmini Wickramasinghe
Illinois Institute of Technology, USA

Eliezer Geisler
Illinois Institute of Technology, USA

INTRODUCTION

Bionanotechnology is a combination of three terms: "bios" meaning "life," "nano" (origin in Greek) meaning "dwarf," and "*technologia*" (origin in Greek—comprised of "*techne*" meaning "craft" and "*logia*" meaning "saying"), which is a broad term dealing with the use and knowledge of humanity's tools and crafts. Bionanotechnology is a term coined for the area of study where nanotechnology has applications in the field of biology and medical sciences.

Healthcare is defined by the Oxford Dictionary as the "care for the general health of a person, community, etc., especially that is provided by an organized health service." Moreover, *Healthcare* can also be defined as the prevention, treatment, and management of illness and the preservation of mental and physical well-being through the services offered by the medical, nursing, and allied health professions[1].

BACKGROUND

The healthcare industry is one of the world's largest and fastest-growing industries[2]. However, since ages the gravest dilemma of the healthcare system has been its soaring costs. Healthcare forms a major part of a country's economy by consuming over 10 percent (Figure 1) of gross domestic product (GDP) of most developed nations. In 2003, healthcare costs consumed 15.3%[3] of the GDP of the United States, the largest of any country in the world. In 2001, for the Organisation for Economic Cooperation and Development (OECD) countries the average was 8.4%[4] with the United States (13.9%), Switzerland (10.9%), and Germany (10.7%) being the top three (See Appendix A).

Figure 1. U.S. healthcare costs[3]

1950	$12.7 billion	4.5 percent
1965	$40 billion (est.)	6 percent
1980	$230 billion	9 percent
2000	$1.2 trillion	14 percent

At present, the U.S. healthcare system has three prime goals: the provision of high-quality care, ready access to the system, and reasonable costs. Therefore, discovering or inventing new ways which would offer effective and efficient quality healthcare treatment is becoming a main concern, not only in the U.S., but also worldwide, as the expenditure in healthcare is increasing exponentially (Wickramasinghe, Choudhary, & Geisler, in press).

ROLE OF BIONANOTECHNOLOGY IN HEALTHCARE

Bionanotechnology deals with nanomaterials and their applications in life sciences. Bionanotechnology being one of the most promising and growing industry today, a large amount of work that is being carried out in this area. Bionanotechnology is currently in its prime with researches being carried out all over the globe in this field. The applications[5,6] of bionanotechnology are vast. The following outlines what is currently being done in this field:

- **Dendrimers:** (Svenson & Tomalia, 2005) It can act as a good *drug delivery agent* as well as a *time-release delivery agent*. It is normally used in therapeutic developments which are targeted to cancer cells (Baker, 2003). Dendritic polymers are also used for the development of nanosensor/NEMS systems for noninvasive, continuous monitoring of astronauts for the biologic effects of space travel.

- **Gold nanoshells:** They are nanoparticles made of silica (glass) which are completely coated with gold (Ha, Jeong, & Chung , 2005; Jain, 2005). Researchers are trying to make use of gold nanoshells for the detection as well as treatment of cancerous cells.
- **Patient-specific medicine:** (Baba, 2006) Personalized medicine or Patient-Specific medicine simply means the *prescription of specific therapeutics best suited for an individual*. Xiaolian Gao, professor of chemistry, has developed a chemical process for building a device that could help doctors predict a patient's response to drugs or screen patients for thousands of genetic mutations and diseases, all with one simple lab test—on a DNA chip. The ultimate vision of this research is the development of Patient-specific medicine.
- **BioNanosensors[7]:** Researchers in many universities and R&D laboratories are trying to develop sensors based on nanotechnology which, when implanted under the skin of a human being, will be able to detect the level of glucose, hormones or cholesterol. Nanosensors (Clark, Singer, Korns, & Smith, 2004; Jain, 2005; Kohli & Martin, 2005; Kvennefors & Persson, 2004) can play a vital role in the treatment of genetically-based diseases like sickle cell anemia, due to their characteristics of have high sensitivity and specificity.
- **Smart drugs[8,9]:** Drugs when injected inside the human body will be able to predict the specific location where the human cells have been contaminated, and then release a precisely targeted drug dose (Peppas et al., 2004).
- Biochemist Susan Hardin and four University of Houston colleagues are developing a new technology for direct molecular sensing that could be used to sequence an individual's entire genome—the gathering of all the genetic information contained in a person's DNA—in less than 24 hours, rather than 2–3 days.
- **Quantum dots[10]:** "Quantum dots" are nanoparticles made out of semiconductors having unique electrical and optical properties (Jiang, Papa, Fischer, Mardyani, & Chan, 2004; Jain 2005; Chan, 2006; Weng & Ren, 2006). They are specifically used for *imaging purpose* inside the human body. They demonstrate extremely high sensitivity.
- **Silver nanoparticles:** As the term suggests, they are nanoparticles derived from the silver atom. A study conducted by the University of Texas and Mexico University found that silver nanoparticles can kill HIV-1, and is likely to kill virtually any other virus.
- **Regenerative medicines:** (Emerich & Thanos, 2003; Rajangam, Behanna, Hui, Han, Hulvat, Lomasney, & Stupp, 2006; Yamato & Okano, 2006) Regenerative medicine helps natural healing processes to work faster, or uses special materials to regrow missing or damaged tissue. Scientists at Northwestern University are working towards the development of regenerative medicines. They have successfully *grown nerve cells* using an artificial three-dimensional network of nanofibers. They have also developed a liquid that forms a gel-like mass of nanofibres on contact with water, which could provide the most promising vehicle yet for the *regeneration of damaged spinal cords*. Many scientists working in the area of regenerative medicine are trying to develop *artificial scaffolds* that store or attract cells, and then control their growth and final identity.

With such great applications, bionanotechnology can help in the reduction of the soaring costs of healthcare. It can help the healthcare system in the production of better and smarter drugs. Expenditures can be highly reduced along with the improvement in the quality of healthcare by the inventions of smaller devices for healthcare purposes and better surgical interventions. Moreover, with faster curing time, the number of days a patient would have to stay in a hospital will lessen significantly, resulting in lower costs of healthcare.

Although most bionano products are currently under research or production, there number has been increasing exponentially in the market. Some of the bionano products that are currently available in the market are *Abraxane* developed by Abraxis Bioscience[11], *M-DNA* developed by Advance Technologies[12], *PuraMatrix*[13] is 3DM Inc.'s patented product line of synthetic self-assembling peptide hydrogels invented at MIT, *Opticell* developed by Bio Crystal Ltd.[14], *Nanosensors and NanoBiosensors* being developed by Bionano International Singapore Pte Ltd.[15], *The Verigene™ System* by Nanosphere[16], *Quantum Dots* by Evident[17], *Nanodisc*[18] developed by Nanodisc Inc., and many more. With such bionano products in the market, and more to be made readily available with the next three to five years of time frame, bionanotechnology shows a tremendous potential in lowering down the costs of healthcare system.

BIONANOTECHNOLOGY AND SOCIETY

In terms of bionanotechnology, the society has been divided into two major categories: one fraction of the society considers that bionanotechnology can prove to be a great benefit to mankind, while the other part raises issues and concerns relating to safety, moral, social, legal, and ethical behavior. Some of the major concerns[19] related to bionanotechnology are discussed below.

- *Will we be able to differentiate at nanoscale?* This is a very important issue raised by the advent of nanotechnology. How will we be able to distinguish between an infected cell and a good cell in our body at nanoscale where everything looks the same? According to Dr. Kevin FitzGerald:

  "We cannot easily define what a disease is at nanoscale. The same genetic sequence may be considered a mutation in one context and merely a difference or polymorphism in another context[…]Bionanotechnology is so reductionistic in this sense that one can even lose the biological concepts of health and disease when focusing on this scale alone."

- *Once in the body, can nanoparticles evade natural defenses of humans and other animals?* What is the likelihood of immune system recognition of nanomaterials? According to Dr. William Hurlbut, nanoparticles might have a subtle effect on the body that we should try to be attentive to. They might interact with basic cell mechanisms, block up pores or passages and change the mechanical forces that influence cell differentiation. Alterations at such a fundamental level of growth and development could have adverse consequences that remain unrecognized for years.
- *Where do we stop?* (Connelly, 2005) It has been brilliantly described by Dr. E. Geisler in his book—*Creating Values With Science & Technology*[20]—that once the Pandora's Box has been opened, it becomes impossible to put the evils back inside the box, meaning that it is impossible to reverse the advancements in the field of science and technology, and so is the case with bionanotechnology. Hence, we need to take care of what we are developing else the consequences can be disastrous.
- *Will genetic modifications that increase sustainability breech our current definition of being "human?"*[21] Humans by nature are supposed to be mortals while machines are considered to be immortals. Thus, by using devices and procedures to indefinitely lengthen human life are we not going to turn humans into machines? Dr. Nilmini Wickramasinghe thinks that Bionanotechnology can surmount the nature's rule of "survival of the fittest in the animal world." She gave an example of a premature baby who, according to nature, is suppose to perish, but because of various medical facilities, can easily survive. This is neither good for the baby nor for the nature's phenomena.
- How will liability and criminal statuses apply to the misuse of nanotechnology innovations?
- *Will it be just another way for rich nations to get richer, or should we take steps now to ensure the developing and underdeveloped countries can participate?* The technology required to genetically "improve" a human being is not going to be inexpensive. According to Dr. FitzGerald, as this technology will be very expensive, certainly at its inception, it will be primarily available to the wealthy to use it to their own benefit. In this way the gap between the "have"s and the "have-not"s will become even greater. So, *is bionanotechnology going to be the tool of the rich?* Also, there is already a wide divide between the various nations of the world, thus it can be an ethical concern to prevent the global "nanodivide" from further contributing to uneven power and wealth distribution.
- *Questions relating to privacy of an individual.* Imagine tiny sensors traveling through your bloodstream, sending constant signals to nearby

computers which relay information about you to your doctor, attorney, local fire or police station, grocer, creditor, or financial institution.[22]

- At present, very little is known about the *health and environmental effects of bionanotechnology* (Preston, 2005). The eventual proliferation and use of nanotechnology could cause unintended consequences, such as the creation of new classes of toxins or related environmental hazards[23]. According to Mr. Sourabh Bansal, MS student at Illinois Institute of Technology:

  "Regular use of bionanotechnology esp. bionanorobots in humans can decrease the immunity of our coming generations towards diseases as their cure will become highly dependent on bionanotechnology. The count of white blood cells in a human body will most likely extinct gradually due to the excessive use of bionanotechnology for curing purposes."

- Who should define these as needs, or judge whether the expense of research is better spent on other ends?

Moreover, some great scholars also believe that bionanotechnology is being greatly hyped. Among them is Dr. Michele Mekel, Associate Director at Institute on Biotechnology & the Human Future, IIT Chicago. According to her, *"nanobiotechnology's biggest problem is 'hype'"*. She believes, *"It has been hyped so much that it may fail. People are destined to be disappointed."* Moreover, Dr. Mekel also said, *"Companies are trying to capitalize 'nano' as a brand."* However, according to Dr. FitzGerald:

Nanotechnology is being hyped because the hype helps to generate more funding. It is good to inform the public about this hype because this situation raises important fundamental questions about the relationship between science and technology development and the public trust.

In addition, Mr. George Kimbrell, Staff Attorney at the Intenational Center for Technology Assessment and the Center for Food Safety, considers nanotechnology as a *"New Wonder Technology."* He said:

There has been cultural amnesia about stuff in the past, this is not the first time. Every time we discover something new it is always hyped[…]This area of nanotechnology is the least researched and regulated. There has been no cost benefit analysis and a significant amount of hype.

In spite of such intense arguments concerning the consequences of bionanotechnology, there is still another section of the society which believes that, if used wisely, bionanotechnology can open doors to great inventions and discoveries, which can help not only reduce the costs of healthcare, but also highly improve the quality of life of a human being. One such scholar who believes that bionanotechnology can prove to be a boon to mankind is Mr. Nik Rokop, President and CEO, nLake Technology Partners, LLC. According to him, *"Historically, people are skeptical of changes."* He stated a brilliant example of nuclear technology. When it was being discovered, there was a huge concern about its affects. However, as we see now nuclear science provides us with nuclear energy as the unlimited source of energy which is truly remarkable. According to Mr. Rokop, people do not really know much about what is really happening in the field of nanobiotechnology, and are hence making assumptions. He said, *"Discussions should be based on facts and not emotions."* When asked if it is correct to mutate a human being he promptly replied, *"Nature creates mutation in the form of evolution which is created due to the presence of certain external factors. A world with no technological progress is nothing but Stone Age. I would not like to be a part of such world."*

Dr. Jialing Xiang, assistant professor of Biology, IIT Chicago, also considers nanobiotechnology as a benefit to human society, due to its focus on detection and increase in sensitivity. She said, *"There is no problem if this technology is solely used for nano level detection. It is a very powerful tool, like an electronic microscope, which can give the minutest of the details inside a human being and help in the early and effective diagnosis of a disease."*

According to Dr. Ruthanna Gordon, assistant professor, Institute of Psychology, IIT Chicago, *"claiming all positive or negative effects of a particular technology is silly."* Hence, we need to understand that there is *nothing intrinsically good or bad about a technology, it all depends on how it is used*.

CONCLUSION

Nanotechnology shows a tremendous potential for exceptional works in the area of medicine and biotechnology. This area of nanotechnology which deals with biology and life sciences is also known as bionanotechnology. In future, bionanotechnology can help in the early and effective diagnosis of various diseases like cancer, Alzheimer's, and so on.

We discussed some of the most prominent and researched applications of bionanotechnology like dendimers, personalized medicines, bionanosensors, regenerative medicines, quantum dots, gold nanoshells, silver nanoparticles, smart drugs, and so on. Moreover, we also mentioned a number of products that are currently available in the market based on bionanotechnology. In addition, as discussed above bionanotechnology can also prove to be an excellent tool for the reduction in cost of healthcare industry as it can provide us with better and effective drugs, smaller devices, and better surgical interventions.

Discussions above also describe the views of the two divisions of the society in case of bionanotechnology. The ethical, social, moral, and legal concerns and issues of bionanotechnmology are also discussed, along with valuable opinions from various prominent scholars across the United States.

ACKNOWLEDGMENT

The authors would like to acknowledge the contributions of George Kimbrell (International Center for Technology Assessment[25]), Jialing Xiang (Assistant Professor, Illinois Institute of Technology), Kevin FitzGerald (Georgetown University), Michele Mekel (Institute on Biotechnology & the Human Future), Nik Rokop (nLake Technology Partners, LLC), Ruthanna Gordon (Assistant Professor, Illinois Institute Of Technology), Sourabh Bansal (MS Student, Illinois Institute of Technology), William B. Hurlbut (Stanford University Medical Center).

REFERENCES

Baba, Y. (2006a). Nanotechnology in medicine. *Nippon Rinsho*, *64*(2), 189–198.

Baba, Y. (2006b). Diagnosis of diseases by nanodevice. *Nippon Rinsho*, *64*(2), 271–277.

Bennett, D.J., & Schuurbiers, D. (2005). Nanobiotechnology: Responsible action on issues in society and ethics. *Conference and Trade Show, European Federation of Biotechnology* (Vol. 2, pp. 765–768).

Blackwelder, B. *Exploring the risks and ethics of nanotechnology*. Retrieved February12, 2008, from http://www.nanoand-society.org/commentaries/commentary_blackwelder01.html

Cameron, N.M. de S. *Nano, ethics and society: lessons from the European "Frankenfood" debacle*.

Chan, W.C. (2006). Bionanotechnology progress and advances. *Biol Blood Marrow Transplant*, *12*(1), 87–91. Suppl 1.

Clark, J., Singer, E.M., Korns, D.R., & Smith, S.S. (2004). Design and analysis of nanoscale bioassemblies. *Biotechniques*, *36*(6), 992–996, 998–1001.

Connelly, M.A. (2005). Nanotechnology – its promises for the field of medicine and the ethical implications of these promises. *News from the bottom, 2005*.

Emerich, D.F. (2005). Nanomedicine--prospective therapeutic and diagnostic applications. *Expert Opin Biol Ther*, *5*(1), 1–5.

Emerich, D.F., & Thanos, C.G. (2003). Nanotechnology and medicine. *Expert Opin. Biol. Ther.*, *3*(4), 655–663.

Geisler, E. (2001). *Creating value with science and technology* (pp. 282–315). Westport, CT: Greenwood Publishing Group. ISBN 1567204058.

Gordon, R. (2005, December). Future hopes, future fears: Predicting public attitudes toward nanotechnology. *Nano and Society Newsletter*.

Grodzinski, P., Silver, M., & Molnar, L.K. (2006). Nanotechnology for cancer diagnostics: Promises and challenges. *Expert Rev Mol Diagn*, *6*(3), 307–318.

Ha, T.H., Jeong, J.Y., & Chung, B.H. (2005). Immobilization of hexa-arginine tagged esterase onto carboxylated gold nanoparticles. *Chem Commun (Camb)*, *31*, pp. 3959–3961. Retrieved February 12, 2008, from http://www.rsc.org/ej/CC/2005/b504184h.pdf

Jain, K.K. (2005). Nanotechnology in clinical laboratory diagnostics. *Clin Chim Acta*, *358*(1–2), 37–54.

Jiang, W., Papa, E., Fischer, H., Mardyani, S., & Chan, W.C. (2004). Semiconductor quantum dots as contrast agents for whole animal imaging. *Trends Biotechnol.*, *22*(12), 607–609.

Kohli, P., & Martin, C.R. (2005). Smart nanotubes for biotechnology. *Curr Pharm Biotechnol, 6*(1), 35–47.

Kvennefors, A., & Persson, F. (2004). *The technology and applications of bionanosensors*. Retrieved February 12, 2008, from http://nanobio.ftf.lth.se/~biokurs/projects/projects2004/7C_NanoBioSensor.pdf

Lin, H., & Datar, R.H. (2006). Medical applications of nanotechnology. *Natl Med J India, 19*(1), 27–32.

Mekel, M. (2006a). A review of nanohype: The truth behind the nanotechnology buzz. *Nanotechnology Law and Business, 3*(3).

Mekel, M. (2006b). Nanotechnologies: Small science, big potential and bigger issues. *Ingentaconnect*, *49*(4), 47–53.

Miller, S.E. (2003). Converging technologies, emerging challenges: Societal, ethical and legal issues. *Seminar at the U.S. Department of Commerce*. Retrieved February 12, 2008, from http://www.ctba.us/articles/CTBA.DOC.110603.Summary.SEM-1.doc

Peppas, N.A. (2006). *Smart drug delivery and bionanotechnology*. Retrieved 12, 2008, from http://www.che.utexas.edu/research/biomat/PDFReprints/PeppParma05.pdf

Preston, C.J. (2005). The promise and threat of nanotechnology can environmental ethics guide U.S.? *HYLE--International Journal for Philosophy of Chemistry*, *11*(1), 19–44.

Rajangam, K., Behanna, H.A., Hui, M.J., Han, X., Hulvat, J.F., Lomasney, J.W., & Stupp, S.I. (2006). Heparin binding nanostructures to promote growth of blood vessels. *Nano Lett, 6*(9), 2086–2090.

Svenson, S., & Tomalia, D.A. (2005). Dendrimers in biomedical applications--reflections on the field. *Adv Drug Deliv Rev.*, *57*(15), 2106–2129.

Weng, J., & Ren, J. (2006). Luminescent quantum dots: a very attractive and promising tool in biomedicine. *Cur Med Chem.*, *13*(8), 897–909.

Wickramasinghe, N. Patient Centric Health Information System in the U.S. In M. Quigley (Ed.), *Encyclopedia of Information Ethics and Security.* Herhsey, PA: Idea Group Publishing.

Wickramasinghe, N., Choudhary, S., & Geisler, E. (in press). Bionanotechnology: Its applications and relevance to healthcare. *International Journal Biomedical Engineering and Technology.* InterScience.

Yamato, M., & Okano, T. (2006). Nanotechnology-based regenerative medicine--cell sheet engineering utilizing temperature-responsive culture dishes. *Nippon Rinsho*, *64*(2), 335–341.

KEY TERMS

Drug Delivery Agent: Drug delivery agent is a medium which facilitates in the delivery of a pharmaceutical compound to humans or animals.

Healthcare Costs: Cost paid to hospitals, physicians, nursing homes, diagnostic laboratories, pharmacies, medical device manufacturers, and other components of the healthcare system.

Nanosensors: Nanosensors can detect either minute particles or miniscule quantities of something.

Nanobiosensors: Nanobiosensors are capable of penetration and location at specific sites within single living cells, which is made possible through developments in nanobiotechnology.

Nanoshell: A spherical core consisting of a particular compound, which is surrounded by a shell of a few nanometer of thickness.

Nanoparticles: A microscopic particle whose size is measured in nanometres (nm). It is defined as a particle with at least one dimension <100nm.

Society: An organized group of persons associated together for religious, benevolent, cultural, scientific, political, patriotic, or other purposes.

ENDNOTES

1 http://www.answers.com/healthcare&r=67#copyright
2 http://en.wikipedia.org/wiki/Healthcare
3 http://www.cms.hhs.gov/statistics/nhe/historical/highlights.asp
4 http://www.oecd.org/dataoecd/10/20/2789777.pdf
5 http://www.researchandmarkets.com/reports/238673
6 http://inbt.jhu.edu/index.php
7 http://www.nanorobotdesign.com/
8 http://www.pro-pharmaceuticals.com/aboutus.htm
9 http://www.ibnam.northwestern.edu/
10 http://www.lanl.gov/mst/nano/computing.html
11 http://www.abraxisbio.com/
12 www.advancetechnologies.net/
13 http://www.puramatrix.com/
14 http://www.biocrystal.com/
15 http://www.bionano.com.sg/index.html
16 http://www.nanosphere-inc.com/
17 http://www.evidenttech.com/
18 http://www.nanodiscinc.com/technology.htm
19 http://www.nanoand-society.org/
20 Creating Values With Science & Technology, Eliezer Geisler; ISBN: 1567204058; Publication of Greenwood Publishing Group, Inc.; First Published in 2001; Pg: 282-303
21 Nanotechnology – its promises for the field of medicine and the ethical implications of these promises By Marie A. Connelly
22 Converging technologies, emerging challenges: societal, ethical and legal issues by: Sonia e. Miller Seminar at the U.S. department of commerce November 06, 2003
23 Zhang and Masciangioli, 2002
24 http://www.nanoand-society.org/about/nano_101_impact.html
25 http://www.icta.org/template/index.cfm
26 http://www.oecd.org/dataoecd/10/20/2789777.pdf

APPENDIX A[26]

Table 1. Growth of expenditure on health, 1990-2001

	Real annual per capita growth rates, 1990-2001 (%)(1)		Health spending as percentage of GDP(2)		
	Health spending	GDP	1990	2000	2001
Australia	3.8	2.4	7.8	8.9	..
Austria	2.6	1.8	7.1	7.7	7.7
Belgium	3.5	1.7	7.4	8.6	9.0
Canada	2.3	1.6	9.0	9.2	9.7
Czech Republic	4.1	0.6	5.0	7.1	7.3
Denmark	1.9	1.9	8.5	8.3	8.6
Finland	0.5	1.6	7.8	6.7	7.0
France	2.5	1.5	8.6	9.3	9.5
Germany	2.0	1.2	8.5	10.6	10.7
Greece	4.0	1.8	7.4	9.4	9.4
Hungary	2.1	2.6	7.1	6.7	6.8
Iceland	2.8	1.6	8.0	9.3	9.2
Ireland	6.7	6.2	6.1	6.4	6.5
Italy	1.9	1.4	8.0	8.2	8.4
Japan	3.8	1.1	5.9	7.6	..
Korea	7.4	5.2	4.8	5.9	..
Luxembourg	3.0	3.9	6.1	5.6	..
Mexico	4.9	1.4	4.5	5.6	6.6
Netherlands	3.1	2.1	8.0	8.6	8.9
New Zealand	3.0	1.5	6.9	8.0	8.1
Norway	3.5	2.8	7.7	7.7	8.3
Poland	5.0	3.3	5.3	6.0	6.3
Portugal	6.1	2.3	6.2	9.0	9.2
Slovak Republic	..	..	..	5.7	5.7
Spain	3.4	2.3	6.7	7.5	7.5
Sweden	2.1	1.5	8.2	8.4	8.7
Switzerland	2.4	0.2	8.5	10.7	10.9
United Kingdom	4.2	2.0	6.0	7.3	7.6
United States	3.2	1.7	11.9	13.1	13.9
OECD average (3)	3.4	2.1	7.3	8.1	8.4

Notes:
(1) Australia, Japan, Korea, Luxembourg 1990-2000, Hungary 1991-2001, Germany 1992-2001.
(2) Hungary 1991, Germany 1992.
(3) OECD average excludes Slovak Republic and Turkey. The OECD average for 2001 includes data for 2000 for Australia, Korea, Japan and Luxembourg.
Growth in health spending and GDP are calculated based on 1995 GDP constant prices.
No recent estimates are available for Turkey.

Source OECD Health Data 2003.

C

Critical Success Factors for Delivering M-Health Excellence

Nilmini Wickramasinghe
Illinois Institute of Technology, USA

Steve Goldberg
INET International Inc., Canada

INTRODUCTION

Medical science has made revolutionary changes in the past decades. Contemporaneously, however, health care has made incremental changes at best. The growing discrepancy between the revolutionary changes in medicine and the minimal changes in health care processes is leading to inefficient and ineffective health care deliver and one, if not the most significant, contributor to the exponentially increasing costs plaguing health care globally.

Health care organizations can respond to these challenges by focusing on three key solution strategies (or the value proposition): (1) access—caring for anyone, anytime, anywhere; (2) quality—offering world-class care and establishing integrated information repositories; and (3) value—providing effective and efficient health care delivery. These three components are interconnected such that they continually impact the other, and all are necessary to meet the key challenges facing health care organizations today.

The application of mobile commerce to health care (i.e., m-health) appears to offer a way for health care delivery to revolutionize itself. However, little if anything has been written regarding how to achieve excellence in m-health. This chapter serves to address this major void by presenting an integrative framework for achieving m-health, developed through the analysis of longitudinal applied research conducted by INET in conjunction with academia. After presenting this framework and discussing its key inputs, we then illustrate how the mapping of case data to the model enables the attainment of a successful m-health application to ensue and the benefits of adopting such a methodology.

BACKGROUND

Currently, the health care industry in the United States as well as globally is contending with relentless pressures to lower costs, while maintaining and increasing the quality of service in a challenging environment (Blair, 2004; European Institute of Medicine, 2003; Frost & Sullivan, 2004; Kulkami & Nathanson, 2005; Kyprianou, 2005; Lacroix, 1999; Lee, Albright, Alkasab, Damassa, Wang & Eaton, 2003; National Coalition on Healthcare, 2004; National Center for Health Statistics, 2002; Organisation for Economic Co-operation and Development (OECD), 2004; Pallarito, 1996; Plunkett, 2005; Russo, 2000; Wickramasinghe & Silvers, 2003; World Health Organization Report, 2000, 2004). It is useful to think of the major challenges facing today's health care organizations in terms of the categories of demographics, technology, and finance. Demographic challenges are reflected by longer life expectancy and an aging population; technology challenges include incorporating advances that keep people younger and healthier; and finance challenges are exacerbated by the escalating costs of treating everyone with the latest technologies. Health care organizations can respond to these challenges by focusing on three key solution strategies: (1) access—caring for anyone, anytime, anywhere; (2) quality—offering world-class care and establishing integrated information repositories; and (3) value—providing effective and efficient health care delivery. These three components are interconnected such that they continually impact the other, and all are necessary to meet the key challenges facing health care organizations today.

In short, the health care industry is finding itself in a state of turbulence and flux (European Institute of Medicine, 2003; National Coalition on Healthcare, 2004; Pallarito, 1996; Wickramasinghe & Mills, 2001; World Health Organization Report, 2000, 2004). Such an environment is definitely well suited for a paradigm shift with respect to health care delivery (von Lubitz & Wickramasinghe, 2005). Many experts within the health care field agree that m-health appears to offer solutions for health care delivery and management that serve to maximize the value proposition for health care. However, to date, little if anything has been written regarding how to achieve excellence in m-health. This chapter serves to address this apparent void.

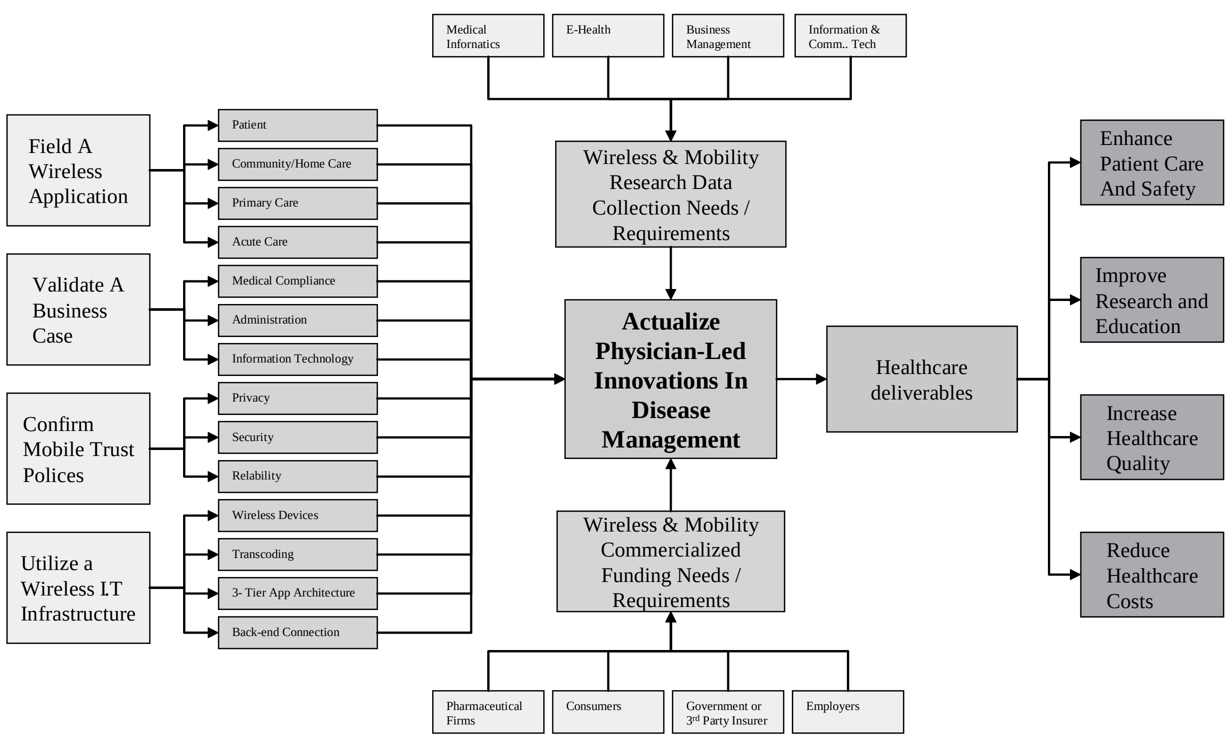
INTEGRATIVE MODEL FORM-HEALTH

Successful m-health projects require consideration of many components. Figure 1 provides an integrative model that serves to capture all the key factors we have identified through our research. What is highlighted by this figure is the need to address many aspects in order to achieve m-health excellence upon which we now expand.

Web of Players

The first consideration is the people component. Any health care initiative, be it wired or wireless, must first and foremost be aware of all key actors involved in the delivery of health care. It is useful to think of these various actors as a web of players, because they interact at different levels and degrees, depending on the specific action or procedure. Figure 2 depicts the web of players that must be considered for health care in general and m-health in particular. From this figure, it is possible to see that m-health requires input and coordination between and within suppliers, payers, health care organizations, providers, regulators, and the patient, if excellence is to truly ensue. Further, all these players are represented in Figure 1, where their specific roles regarding any m-health project are also

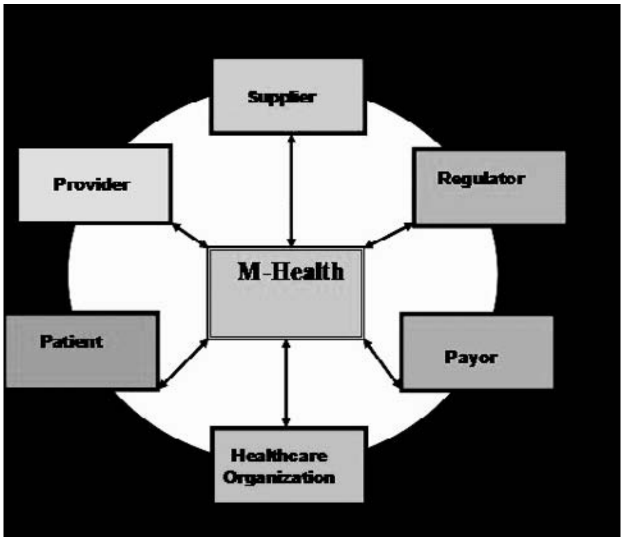
Figure 1. A mobile e-health project delivery model



* Information and Communication Technology

© Goldberg & Wickramasinghe, 2004

Figure 2. Web of healthcare players (Adapted from Wickramasinghe, et al., 2005)



noted. For example, patients, providers, and health care organizations are among the primary inputs, while payers and regulators as well as various suppliers are modifiers to the process of actualizing the physician-led solution.

IT Architecture

The next important consideration is concerned with the existing IT infrastructure and architecture. Any m-health solution must leverage off existing IT architecture, wired and wireless, at the respective locations of the web of key players. Typically, in today's technocentric world, this involves understanding the client-server computing paradigm as depicted in Figure 3. To support such a client-server architecture, special attention must be paid to the ICT infrastructure. The ICT infrastructure includes phone lines; fiber trunks and submarine cables; T1, T3, and OC-xx; ISDN, DSL, and other high-speed services used by businesses as well as satellites; earth stations; and teleports. A sound technical infrastructure is an essential ingredient to the undertaking of e-health and m-health initiatives. Such infrastructures should also include telecommunications, electricity, access to computers, number of Internet hosts, number of ISPs (Internet Service Providers) and available bandwidth and broadband access. Such IT infrastructure and architecture components form a key input into the designing and development of any m-health project, as can be seen in Figure 1.

IT Architecture and Standard Mobile Environment

By adopting a mobile/wireless health care delivery solution, it is possible to achieve rapid health care delivery improvements, which impact both the costs and the quality of health care delivery. This is achieved by using an e-business acceleration project that provides hospitals a way to achieve desired results within a standardized mobile Internet (wireless) environment. Integral to such an accelerated project is the ability to build on the existing infrastructure of the hospital. This then leads to what we call the three-tier Web-based architecture (Figure 4).

In such an environment, Tier-1 is essentially the presentation layer, which contains the Web browser, but no patient data are stored within this layer, thereby ensuring compliance with international security standards/policies such as HIPAA. Tier-2 then, provides the business logic, including but not limited to lab, radiology, and clinical transcription applications; Messaging of HL7, XML, DICOM, and other data protocols and interface engines to a Hospital Information Systems (HIS); Lab Information Systems (LIS); Radiology Information Systems (RIS); as well as external messaging systems such as Smart Systems for Health (an Ontario Health care IT infrastructure project). Finally, there is the Tier-3 architecture, which consists of back-end databases like Oracle or Sybase.

Figure 3. IT architecture (adapted from Wickramasinghe et al., 2005)

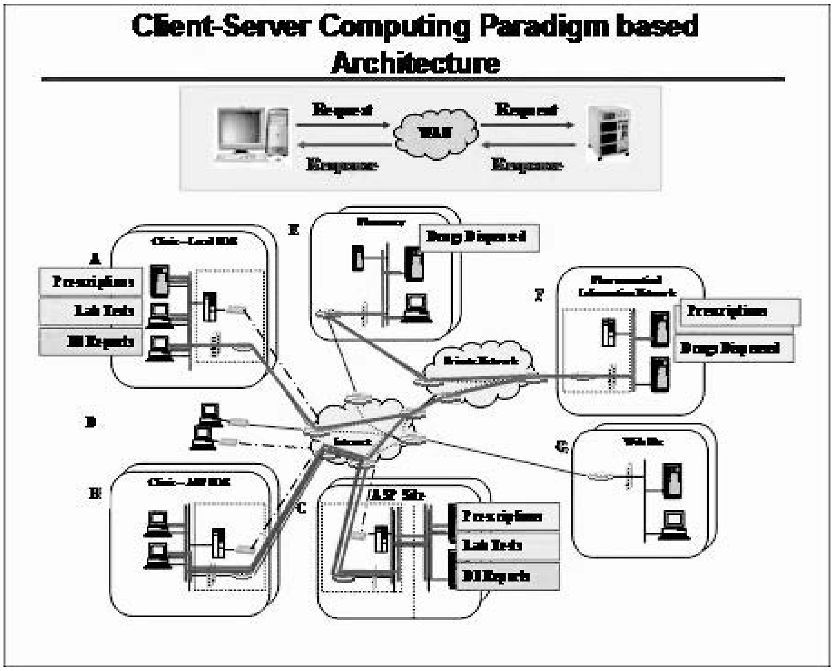
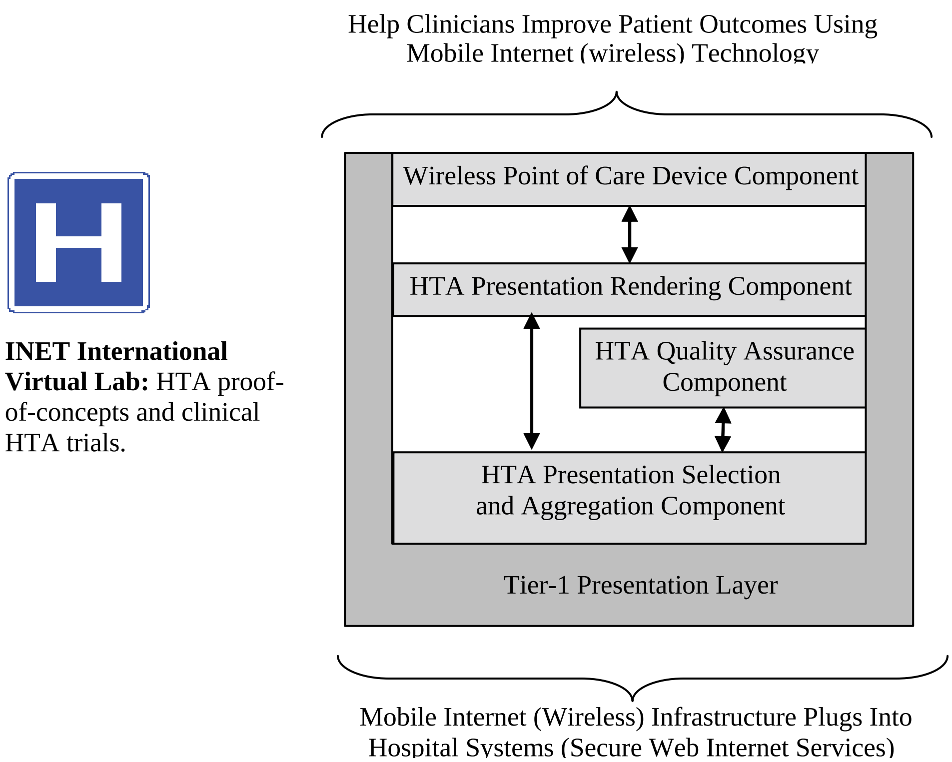


Figure 4. A Standardized Mobile Internet (wireless) Environment



C

Security and Regulatory Conformance

The third primary input in Figure 1 is concerned with security, privacy, and reliability issues. Given the nature of health care data adherence to adequate standards, in this regard it is imperative to the ultimate success of m-health delivery. In the United States, security, privacy, and standards for electronic submissions and exchange of health care information are covered by HIPAA (the Health Insurance, Portability and Accountability Act) (2001) (Fadlalla & Wickramasinghe, 2004; Moore & Wesson, 2002). It is useful to conceptualize this as a HIPAA triangle (Figure 5), which highlights the fundamental elements of the HIPAA regulation; namely, security, transaction standards and privacy.

Security

According to HIPAA, a number of security criteria must be met by all electronic health care transactions. Some of these criteria directly affect how health care systems can be accessed as well as how the key players may interact with these systems. Table 1 details some extracts of the HIPAA security requirements (readers interested in the complete HIPPA security requirements are referred to HIPAA Security requirement matrix, 2002). Essentially, these security criteria fall into three main categories: administrative, physical, and technical. Table 1 summarizes the major issues and levers under each of these categories as well as identifying which are required and which are optional.

Transaction Standards

The Standards for electronic health information transactions cover transactions including claims, enrollment, eligibility, payment, and coordination of benefits. Succinctly stated, the aspect of HIPAA referring to transaction standards can be thought of in terms of practice standards and technical standards, as can be seen in Table 2.

Privacy

The final element of the HIPAA triangle deals with ensuring the **privacy of health care information.** Specifically, the Federal Register (Vol. 67, No. 157) details all the rules that must be adhered to with respect to privacy. The purpose of these rules is to maintain strong protections for the privacy of individually identifiable health information, addressing the unintended negative effects of the privacy requirements on health care quality or access to health care, and relieving unintended administrative burdens created by the privacy requirements. Thus, theses privacy requirements cover uses and disclosures of treatment and payment information and create national standards to protect individuals' medical records and other personal health information. Specifically, they do the following:

- Give patients more *control* over their health information.

Figure 5. HIPAA triangle (Adapted from Fadlalla & Wickramasinghe, 2004)

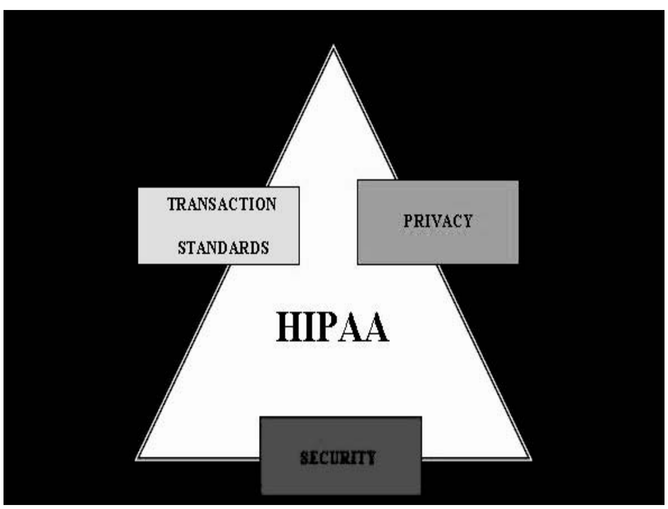


Table 2. Practice and technical standards

STANDARD SETS
Practice Standards
Health Care Common Procedure Coding System (HCPCS) This standard contains the Level II alphanumeric HCPCS procedure and modifier codes, their long and short descriptions. These codes, which are established by CMS's Alpha-Numeric Editorial Panel, primarily represent items and supplies and non-physician services not covered by the American Medical Association's CPT-4 codes. This standard does not contain the American Medical Association's CPT-4 codes.
ICD-9 – Diagnosis Codes International Classification of Diseases, 9th revision, Clinical Modification ICD-9-CM.
ICD-9 – Procedure Codes International Classification of Diseases, 9th revision, Clinical Modification ICD-9-CM.
Technical Standards
Technical Standards Adoption of electronic data interchange (EDI) using health care industry implementation guidelines and other standards such as XML and X12

Table 1. HIPAA security requirements (Adapted from Fadlalla & Wickramasinghe, 2004)

Extracts from HIPAA Concerning Security Requirements
• Establishment of trust partnership agreements with all business partners
• Formal mechanisms for accessing electronic health records
• Procedures and policies to control access of information
• Maintaining records authorizing access to the system
• Assuring that system users receive security awareness training and the training procedures are periodically reviewed and updated
• Maintaining security configuration, including complete documentation of security plans and procedures, security incident reporting procedures, and incident recovery procedures
• Communication and network control, including maintaining message integrity, authenticity, and privacy. Encryption of messages is also advocated for the open network transmission portion of the message.
• Data authentication to ensure that data is not altered or destroyed in an unauthorized manner

- Set *boundaries* on the use and release of health records.
- Establish appropriate *safeguards* that health care providers and others must achieve to protect the privacy of health information.
- Hold violators *accountable*, with civil and criminal penalties that can be imposed if they violate patients' privacy rights.
- Strike a balance when *public responsibility* requires disclosure of some forms of data; for example, to protect public health.

For patients, this means being able to make informed choices when seeking care and reimbursement for care based on how personal health information may be used. Specifically, privacy requirements do the following:

- Enable patients to find out how their information may be used and what disclosures of their information have been made.
- Generally limit release of information to the minimum reasonably needed for the purpose of the disclosure.

- Give patients the right to examine and obtain a copy of their own health records and request corrections.

For the average health care provider or health plan, the privacy regulations require activities, such as:

- Providing information to patients about their privacy rights and how their information can be used.
- Adopting clear privacy procedures for its practice, hospital, or plan.
- Training employees so they understand the privacy procedures.
- Designating an individual to be responsible for seeing that the privacy procedures are adopted and followed.
- Securing patient records containing individually identifiable health information so they are not readily available to those who do not need them.

A NEW ICT INFRASTRUCTURE TO SUPPORT E-BUSINESS ACCELERATION PROJECTS

INET's challenge was finding an ICT infrastructure to support the delivery of an accelerated e-business project. After five years of investigation and practical experience, INET has defined this new ICT Infrastructure as follows:

1. **Simple and low-cost technology.** Internet applications, wireless technology, cell phones, and/or personal digital assistants (PDAs)
2. **The next generation ICT professional.** Engage ICT players that have made the transition to a new cultural of meeting the highly responsive and evolutionary needs of end-users
3. **Demonstrable processes to accelerate ICT project delivery.** Demonstrate a rigorous project delivery process that reduces time cycles by 75% with a 50% reduction in project costs

For an in-depth look at INET findings, let us begin with a definition of an information and communication technology (ICT) infrastructure. It can be defined as a set of technologies such as computer hardware, devices, printers, applications, software, wired and wireless networks, and other technology components. The infrastructure is supported by ICT players that research, develop, deliver, service, and sell technologies to enhance information systems operations. These players also deliver new systems using a rigorous systems development life cycle methodology. In health care, an ICT infrastructure can be referred to as e-Health, and typically contains configurations similar to that depicted in Figure 3.

What is different today? In the past, many people resisted the use of technology to exchange and communicate information. It is well documented that they will not use technology if it is too complex or too costly. What is new today is the commercialization and acceptance of wireless technology. Wireless technology eliminates the costs and complexity in communicating and exchanging information at the point of need, regardless of location or time. This is evident by the widespread use of PDAs and cell phones, along with wireless data networks. Today, more than 33% of physicians use PDAs, and by 2005, it is predicted to grow to 50%[1]. In 2003, PDA usage among medical residents was 85%[2]. More than 172 million American consumers use cell phones[3] with the potential to access health care ICT systems anywhere, anytime.

Even with wireless technology removing the cost and usability barriers to enhance communications and exchange information, INET quickly discovered a new barrier when trying to engage the ICT industry in an e-Business Acceleration Project. This is manifesting itself as a user frustration with the responsiveness of the ICT industry to meet their demands. INET believes this is the result of a project engagement gap between the ICT one- to five-year system development cycle practices and users' demands for immediate and quick results. The e-Business Acceleration Project was design to bridge the gap between ICT and the user. From INET's experience, ICT needs a cultural change to meet the highly responsive and evolutionary needs of an e-business acceleration project. The INET experience clearly showed there was no consensus on an ICT infrastructure (people, process, and technology) to support both benefits of wireless technology and rapid project delivery achievements. INET spent four years researching and developing a roadmap to help the ICT industry make this transition. To release a solution and build consensus, INET founded and chaired the first wireless IT committee for the Information Technology

Association of Canada (ITAC). After the inaugural year, the group delivered the Wireless I.T. Infrastructure Procurement Guidelines in Health care[4].

At the same time of the ITAC Wireless IT Committee, INET started to gather evidence on an ICT infrastructure that can support the rapid and concurrent delivery of INET e-business acceleration projects. After a surveillance of Internet usage, INET selected health care market research in 2003. This industry is quickly shifting traditional quantitative research methods, such as face-to-face or telephone interviewing to online surveys using the Internet. The use of the Internet has shortened field time cycles by as much as 75% with the added benefit of reducing costs upwards to 50% when compared with telephone studies. To confirm these findings, INET completed more than 20 market research engagements involving the United States, the United Kingdom, Canada, Japan, Denmark, Italy, Spain, Australia, Germany, and France. To deliver these e-Business Acceleration Projects, INET assembled and used an ICT Infrastructure with the following key attributes:

- The heart of the INET ICT infrastructure is a form generation technology component(s) to create custom online surveys within three days. This an instrument with the programming capability to deliver custom surveys (forms) on demand; field multiple projects at the same time; integrate with many online panel methods; support, at a minimum, all standard survey question types, including single and multi-selects, grid and three-dimensional grid questions, open text/verbatim, and numeric entry; and able to securely integrate video/audio clips and other embedded technology objects.
- The INET hosting environment employs a fully redundant technical platform hosted in a state-of-the-art data center and capable of handling more than 50,000 surveys per hour. INET also assures these platforms are secure, and when required, can conduct a HIPAA-compliant audit.
- The technology hosting servers have redundant connectivity to the Internet and are managed on a 7-day/24-hour basis.
- When required, it can connect to wireless networks and devices using transcoding technology components. Transcoding component sits on top of a three-tier application architecture to redirect applications for use by a smaller screen, synchronize data, and support many types of wireless devices from laptops to cell phones (see Figure 5).

Using the INET ICT infrastructure, an INET e-business acceleration project for market researchers begins with a questionnaire written in an Microsoft Word document; it is then programmed and hosted for the Internet. When required, the survey is translated into another language to field non-English-speaking countries. Then INET fields the study with an online panel of physicians; for instance, INET sourced almost 200,000 physicians worldwide to participate in online research. An online panel process begins by e-mailing an invitation to the targeted respondent to participate in an Internet survey. With the use of an anonymous electronic ID, privacy is upheld and enhanced. There is no personnel information capture in the online survey. Once a panel member clicks on a unique URL, in the e-mail the member is immediately connected to a secure Internet site and completes the survey. Typically, an online panel response rate is 23% and upwards to 40%. Total time to program and test of a survey is three days. The timing in the field is approximately five to seven days, regardless if this is for N=100 or N=2000 completes, and delivery of the data file is within 24 hours.

In summary, INET reuses the lessons learned from delivering market research data collection projects and using wireless technology to show how an ICT infrastructure can support INET e-business acceleration projects. In this way, continuous improvement is achieved, and the future state is always improved through analysis and diagnosis of the current state outcomes and then implementation of appropriate prescriptive measures. The next step is to gain better acceptance of an e-business acceleration project in health care. This will occur as the industry embraces e-health and wireless possibilities.

ACHIEVING ICT PROJECT SUCCESS IN HEALTH CARE

In order to achieve ICT project success in health care, INET maps the e-business acceleration project to the mobile e-health project delivery model and then vets the model with a use case scenario. INET sees this model, combined with use case scenarios and a peer review process as a rigorous mechanism in the delivery of local and international mobile e-health projects.

For the past five years, the primary sponsors for a mobile e-health project are organizations looking to conduct studies on the use of wireless or mobile technology in health care delivery. These projects are typically funded through government research grants and the IT industry. In INET's case, an e-business acceleration project was used to deliver a mobile e-health project and was funded through the ICT industry sector. However, in 2003, INET began to quickly understand that it would be difficult for the ICT industry to continue the sponsorship of research-originated projects without a much faster commercial payback period. INET clearly saw this as the need to rebalance funding from research to commercial sources. With this in mind, INET started working in collaboration with others on a strategy to incorporate the e-business acceleration project as an INET mobile e-health project and accelerate commercial successes. The INET Mobile e-Health project objectives include the following:

1. Accelerate consensus building with an e-health solution focused on a disease state and driven by the medical model, with the primary objective to streamline communications and information exchange between patients and providers of community/home care, primary care, and acute care.
2. Acquire commercial funding early with a compelling business case. For instance, enhancing therapeutic compliance can improve patient quality of life with significant health care cost savings. It is well documented that in diabetes mellitus, this will have immediate and high impact benefits for health care consumers, pharmaceutical firms, governments, insurers, and employers.
3. Avoid risk by reengineering large-scale health care delivery processes in small manageable pieces. Today, organizations can harness a rigorous method to incrementally enhance a process one step at a time, a way to achieve quick wins early and frequently.
4. Rapid development of simple-to-use, low-cost, private/secure information and communication technology (ICT) solutions. Achieve these benefits through wireless Application Service Providers (ASPs). In addition to rapid development, a wireless ASP can easily connect and bring together many independent health care information systems and technology projects.

To actualize the mobile e-health project, INET is looking to the mobile e-health project delivery model as a framework (Figure 1). For INET, this will support a Mobile e-Health Project Management Office (PMO) to manage the costs, quality, and ICT vendors; deliver many small projects; and replicate projects for local and international distribution. As a first-case scenario for the model, INET is proposing an INET Diabetes Mobile e-Health project with leadership from a family physician. The INET PMO is provisioning a project manager to support this physician-led project to meet both research and commercial sponsors' interests and objectives in diabetes. A detailed description of the key attributes of the INET Diabetes Mobile e-Health project includes the following:

- **Problem statement:** There are many communication and information exchange bottlenecks between patients and their family physicians that prevent the effective treatment of diabetes. As background, a fundamental problem today is the ability to have a private and secure way to manage, search, and retrieve information at the point-of-care. In diabetes, physicians cannot quickly and easily respond to patients with high glucose levels. They need to wait for people to come to the office, respond to phone calls, or reply using traditional mail delivery; they may never receive the patient information.
- **Solution mandate:** Implement a diabetes monitoring program to enhance therapeutic compliance, such as releasing a program to enhance the usage of oral hypoglycemic agents (drugs), and/or the usage of blood sugar monitoring devices.

  As background, everyone wins when enhancing patients' abilities to follow instructions in taking prescribed medication. The patient's health, safety, and quality of life improve with significant health care cost savings. However, it is well documented that many patients do not stay on treatments prescribed by physicians[5]. This is where wireless technology may have the greatest impact to enhance compliance.

  One solution may be as simple as using a cell phone and installing a secure wireless application for patients to monitor glucose levels, and provisioning a physician to use a PDA (personal digital assistant) connected to a wireless network to confidentially access, evaluate, and act on the patient's data.

- **Business case:** In Ontario, Canada, the cost savings may represent almost $1 billion over three years. INET uses a simple calculation to determine the $1 billion savings. This can be found at www.inet-international.com (select the INET mobile e-health project section to review the calculations.

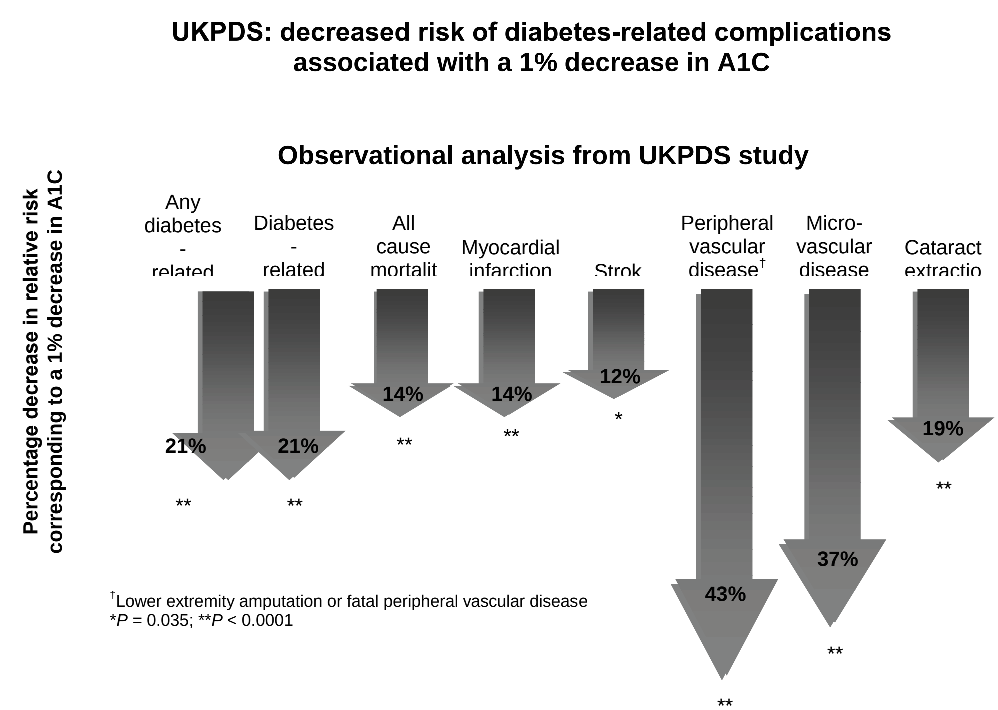
  The business case can be backed with additional data on how the cost of prevention (drugs) is far less than the cost savings associated with reducing the risk of complications associated with diabetes. For instance, the impact of a 1% decrease in A1C is significant and evident in Figure 8.

FUTURE TRENDS

Diabetes is increasing in prevalence globally at alarming rates. We believe the INET solution represents a useful model for effectively enabling the control of this disease so that its sufferers may enjoy a high quality lifestyle. Looking forward, INET's research data indicate that using the Mobile e-Health Project delivery model will increase ICT project success in health care. To realize and test this, INET plans to map the player's form and INET diabetes mobile e-health project (use case scenario) to the model. To show how this may work, review the following mapping exercise. The bold text in black is a project player, and the text in brackets relates to the sections of the model presented in Figure 1.

- **Physician Mobile e-Health Project Lead:** [Actualize physician-led innovations in disease management component in Figure 1]. Physicians provide the linkage to the medical model to enhance disease management programs (wireless Diabetes Program) to enhance patient care and safety, improve research and education, increase health care quality, and reduce health care costs. These are the project deliverables [Health Care Deliverables component in Figure 1].
- **Commercial sponsor(s) [wireless and mobility commercialized fund needs/requirements component in Figure 1]:** The project delivers information and communication solutions for:

Figure 8. Impact form reducing AIC levels



Adapted from Stratton IM, *et al.* UKPDS 35. *BMJ* 2000; 321:405–412.

  - ◦ Consumers wishing to improve their quality of life with an enhanced relationship with their health care providers' (i.e., family) physicians.
  - ◦ Pharmaceutical firms looking to increase revenues with e-compliance programs.
  - ◦ Government/insurers investigating ways to significantly reduce administration and health care costs and shorten health care delivery time cycles (wait times.)
  - ◦ Employers wanting to increase productivity and avoid absenteeism with a healthier workforce.
- **Research sponsor(s) [wireless and mobility research data collection needs/requirements component in Figure 1]:** The project develops intellectual property for researchers in the fields of:
  - ◦ Patient and health care provider relationships
  - ◦ Wireless medical informatics
  - ◦ Therapeutic compliance business case
  - ◦ Wireless information technology usability
- **An INET Mobile eHealth Project delivery team:**
- **Health care delivery team [field a wireless application component in Figure 1]:** For a wireless diabetes program, the players may include:
  - ◦ Health care Consumer: People with Diabetes
  - ◦ Community Care: Nurse Specializing in Diabetes
  - ◦ Primary Care: Family Physician
  - ◦ Acute Care: Endocrinologist
- **Business process analyst [validate a business case component in Figure 1]**
- **Privacy and security consultant [confirm mobile trust policies component in Figure 1]**
- **Programmer using a wireless ASP [utilize a wireless IT infrastructure component in Figure 1]:**
  - ◦ Wireless network and devices
  - ◦ Device and application transcoding
  - ◦ Application service provider
  - ◦ Back-end connection

In conclusion, INET is looking forward to further advancements in the mobile e-health project delivery model to:

- Achieve rapid advancements in health care delivery.
- Improve diabetes management.
- Enhance therapeutic compliance.
- Realize significant health care cost savings[6].

INET is planning to continue its role as a source of use case scenarios for the model with the delivery Mobile e-Health Projects.

CONCLUSION

Health care in the United States and globally is at a crossroads. It is facing numerous challenges in terms of demographics, technology, and finance. The health care industry is responding by trying to address the key areas of access, quality, and value. M-health, or mobile e-health, provides a tremendous opportunity for health care to make the necessary evolutionary steps in order to realize its goals and truly achieve its value proposition. What is important is to ensure m-health excellence. By an in-depth analysis of the rich and longitudinal data of INET, we have developed an integrative macro-level model to facilitate the achievement of m-health excellence. To the best of our knowledge, it is the first such model, and while it is certainly not a panacea, it does help to set the stage and outline the key issues that must be addressed in a systematic fashion so that a successful m-health initiative might ensue. Health care globally must embrace the advances of wired and wireless ICT use to provide superior health care delivery. In such an environment, an imperative becomes successful implementation of these wired and wireless initiatives. The Wi-INET integrative model serves to fill this critical void. We close by strongly urging more research in this area that will further test our framework.

REFERENCES

Blair, J. (2004). Assessing the value of the Internet in health improvement. *Nurs. Times 100,* 28–30.

Boyatzis, R. (1998). *Transforming qualitative information thematic analysis and code development.* Thousand Oaks: Sage Publications.

Crossing the quality chasm—A new health system for the 21st century. (2001). Washington, DC: Committee on Quality of Healthcare in America Institute of Medicine, National Academy Press.

Eisenhardt, K. (1989). Building theories from case study research. *Academy of Management Review, 14*, 532–550.

European Institute of Medicine. (2003). *Health is wealth: Strategic vision for European healthcare at the beginning of the 21st century*. Salzburg, Austria: European Academy of Arts and Sciences.

Fadlalla, A., & Wickramasinghe, N. (2004). An integrative framework for HIPAA-compliant I*IQ healthcare information systems. *International Journal of Health Care Quality Assurance, 17*(2), 65–74.

Frost & Sullivan Country Industry Forecast—European Union Healthcare Industry. Retrieved May 11, 2004, from http://www.news-medical.net/print_article.asp?id=1405

Goldberg, S., et al. (2002). Building the evidence for a standardized mobile Internet (wireless) environment in Ontario, Canada. January 2002 update [internal INET documentation].

Goldberg, S., et al. (2002). HTA presentation rendering component summary [internal INET documentation].

Goldberg, S., et al. (2002). HTA presentational selection and aggregation component summary [internal documentation].

Goldberg, S., et al. (2002). HTA quality assurance component summary [internal INET documentation].

Goldberg, S., et al. (2002). Wireless POC device component summary [internal INET documentation].

Health Insurance Portability and Accountability Act (HIPPA). (2001). Privacy compliance executive summary. Protegrity Inc.

HIPAA security requirement matrix. (2002). Retrieved from- http://www.hipaa.org

Kavale, S. (1996). Interviews: *An introduction to qualitative research interviewing*. Thousand Oaks, CA: Sage Publications.

Kulkarni, R., & Nathanson, L.A. (2005). Medical informatics in medicine. E-medicine. Retrieved from http://www.emedicine.com/emerg/topic879.htm

Kyprianou, M. (2005). Healthcare: Is Europe getting better? In *Proceedings of the European Healthcare Agenda—The European Voice Conference*. EU Health and Consumer Protection. Retrieved from http://www.noticias.info/asp/aspcommunicados.asp?nid=45584

Lacroix, A. (1999). International concerted action on collaboration in telemedicine: G8sub-project 4, *Sted. Health Technol. Inform*., 64, 12–9.

Lee, M.Y., Albright, S.A., Alkasab, T., Damassa, D.A., Wang, P.J., & Eaton, E.K. (2003). Tufts Health Sciences Database: Lessons, issues, and opportunities. *Acad. Med., 78*, 254–624.

Moore, T., & Wesson, R. (2002). Issues regarding wireless patient monitoring within and outside the hospital. In *Proceedings of the 2nd Hospital of the Future Conference*, Chicago.

National Center for Health Statistics: Health Expenditures (2002). Hyattsville, MD: CDC/NCHS. Retrieved from http://www.cdc.gov/nchs/fastats/hexpense.htm

National Coalition on Healthcare. (2004). Building a better health: Specifications for reform. Washington, DC: National Coalition on Healthcare.

Organisation for Economic Co-operation and Development (OECD). (2004). OECD Health Data [CD-ROM]. Retrieved from www.oecd.org/health/healthdata

Pallarito, K. (1996). Virtual healthcare. *Modern Healthcare*, 42–44.

Plunkett's Health Care Industry Almanac. (2005). Houston, TX: Plunkett Research, Ltd.

Russo, H.E. (2000). The Internet: Building knowledge & offering integrated solutions to health care. *Caring, 19*, 18–20, 22–24, 28–31.

Standish Group International Inc. (1994). The CHAOS report. Retrieved from http://www.standishgroup.com/sample_research/chaos_1994_1.php

von Lubitz, D., & Wickramasinghe, N. (2005). Healthcare and technology: The doctrine of networkcentric healthcare [in press]. *Health Affairs*.

Wickramasinghe, N., & Goldberg, S. (2004). How M=EC2 in healthcare. *International J. Mobile Communications, 2*(2), 140–156.

Wickramasinghe, N., & Mills, G. (2001). MARS: The electronic medical record system, the core of the kaiser galaxy. *International Journal Healthcare Technology Management, 3*(5/6), 406–423.

Wickramasinghe, N., & Silvers, J.B. (2003). IS/IT the prescription to enable medical group practices to manage managed care. *Health Care Management Science, 6,* 75–86.

Wickramasinghe, N., et al. (2005). Assessing e-health [forthcoming]. In T. Spil, & R. Schuring (Eds.), *E-health systems diffusion and use: The innovation, the user and the user IT model*. Hershey, PA: Idea Group Publishing.

World Health Organization Report. (2000). *Health systems: Improving performance* (pp. 1–215), Geneva.

World Health Organization Report. (2004). Changing history (pp. 1–167), Geneva.

KEY TERMS

Health Care Value Proposition: The areas that must be addressed to ensure that the health care solution(s) provides value to the patient.

ICT (Information Communication Technology): The combination of computers, hardware, software, networking, and telecommunications technologies.

IT Infrastructure: The technology and networking foundations required to support the m-health solution.

Mobile Health Care: The delivery of health care solutions that incorporates wireless or mobile technology.

Web of Players: The multidimensional group of players required in the delivery of health care solutions.

Current Issues and Future Trends of Clinical Decision Support Systems (CDSS)

Omar F. El-Gayar
Dakota State University, USA

Amit Deokar
Dakota State University, USA

Matthew Wills
Dakota State University, USA

INTRODUCTION

Computer-based clinical decision support systems (CDSSs) leverage developments in information technology to support healthcare providers with clinical decisions as they relate to the improvement of diagnostic, therapeutic, quality and safety processes and outcomes. In the last 50 years there have been significant advances with CDSSs, and the potential for a substantial role in the future of healthcare delivery remains likely. During this time, there have been three main stimulating factors for the development of CDS: (1) the intellectual challenge, (2) the need to address issues relating to patient safety, quality and access to healthcare, and (3) the business and policy reasons relating to the allocation of limited resources and control of spiraling healthcare costs (Greenes, 2006).

The undeniable intellectual challenge to understand and improve upon the human cognitive process has been and continues to be a major driving factor for the development of CDSSs. The human mind has become the model for developing complex systems capable of broadly interactive decision support. At the same time, efforts to reduce medical errors, and increase quality and patient safety have been championed as potential benefits of CDS. Estimates from a 1999 report, by the Institute of Medicine (IOM) indicated that as few as 48,000 and as many as 98,000 people die each year in the U.S., at a cost of more than $6 billion per year, as a result of medical errors (IOM, 1999). Similarly, a recent 2001 IOM report made a passionate plea for closing the quality gap between best practices and actual practices (IOM, 2001).

In this regard, the objective of the article is twofold: (1) to provide a framework for understanding CDSS and within this framework address a number of the current issues facing CDSS, and (2) to suggest likely directions for future research and clinical application.

BACKGROUND

A number of definitions for CDSS exist in the literature. For example, Peleg and Tu (2006) define CDSS as "any computer program to help health professionals make clinical decisions" while Musen, Shahar, and Shortliffe (2001) refer to CDSS as tools for information management, for focusing attention, and for providing patient-specific recommendations. More specific definitions include CDSSs as computer systems designed to impact clinician decision making about individual patients at the point in time when the decisions are made (Berner & La Lande, 2007) and CDSS as computer-based systems for bringing relevant knowledge to bear on the healthcare of the patient (Greenes, 2006). In effect, a CDSS provides patient-specific alerts and reminders, supports differential diagnosis through the interpretation of clinical results and user input, and provides methods to retrieve intelligently filtered knowledge that is medically relevant to the decision making process. Regardless of a particular definition, it is evident that CDSS are *computer-based* systems that leverage *medical knowledge* and *patient-specific data* to respond to a request for decision support by providing *recommendations* with the ultimate goal of improving the quality of the service provided to the *patient*. Figure 1 depicts the key elements in a CDSS. It should be noted however, that the decision support may be requested by a healthcare professional such as a doctor, a nurse, or a technician, another computer system, or the patient. Moreover, while the diagram depicts knowledge and the processing engine as distinct entities, in practice, depending on the underlying technology, the knowledge and the engine may be closely intertwined.

A FRAMEWORK FOR UNDERSTANDING CDSS

In order to synthesize different perspectives on CDSS research, development, and practical issues, we develop a framework (shown in Figure 2) in this section, building on extant literature.

In this framework, CDSS can be characterized along multiple dimensions, such as the type of support and intervention provided by CDSS, and the type of knowledge and methods employed in these systems (Degoulet & Fieschi, 1997). According to Degoulet, there exist two types of support provided by CDSS: (1) support in assessing the patient's state (e.g., in diagnosis or prognosis), and (2) support in suggesting the best course of action (e.g., recommended tests or treatment). Frequently, CDSS in practice may provide both types of support.

CDSS can also differ with respect to the type of intervention ranging from passive systems to active systems (Degoulet & Fieschi, 1997). In passive mode, the user has to initiate the process by providing a request to the system. Such request can be for an advice regarding a particular case, or for feedback (critique) regarding a particular assessment or planned course of action. On the other extreme, active systems can make decisions in an autonomous manner (i.e., independent of the physician or care giver, e.g., dialysis monitors).

CDSS can be further characterized by the type of knowledge or method employed. As noted by Degoulet, Jean, Engelmann, Meinzer, and Jagermann (1995), knowledge utilized for decision support may come from the medical literature encapsulating academic knowledge, and expertise acquired through medical practice. In many cases, systems encapsulate knowledge from both sources. Observations, while not knowledge per se, play a major role in machine learning techniques, and in validating decision models.

A number of methods have evolved over the years as shown in Table 1. Logical methods seek to represent and evaluate logical conditions, and are often used to generate clinical reminders, alerts, and make diagnostic or therapeutic inferences. Research dating back to the late 1950s suggested that the use of these methods as basic building blocks needed to emulate the medical decision-making process (Ledley & Lusted, 1959). Even today, the contemporary clinical system *Retrogram*, founded on the logic-based PRO*forma* application is being widely used in the UK to manage HIV care. Many existing systems in use today typically utilize the evaluation of logical conditions to narrow the range of possibilities as a precursor to applying probabilistic, heuristic, or other methods.

Figure 1. Clinical decision support systems (CDSS)

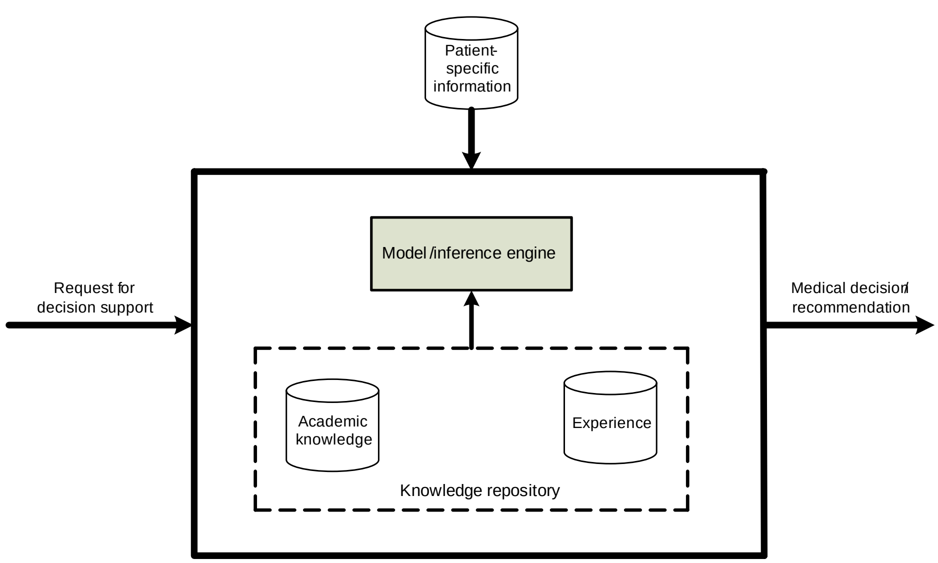


Figure 2. A framework for understanding CDSS

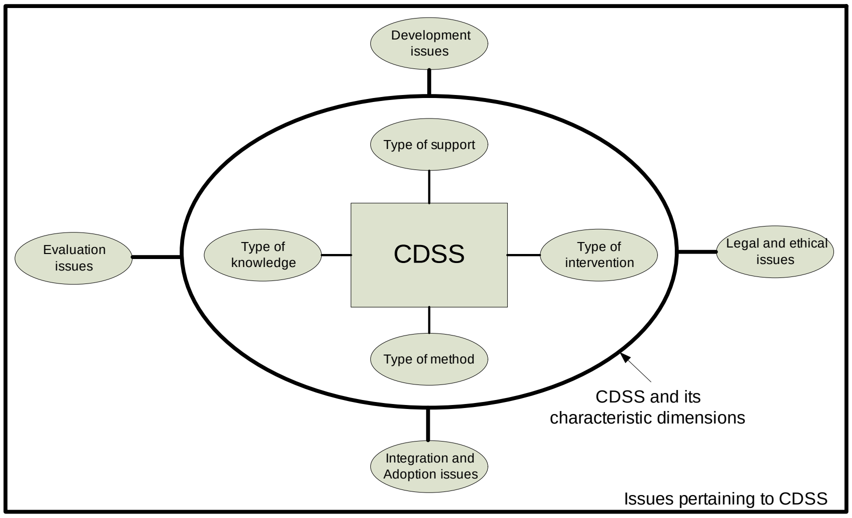


Table 1. Examples of CDSS methods and their uses

CDSS methods	Examples of methods	Examples of uses	Examples of current systems
Logical methods	Decision tables, Venn diagrams, branching logic	Alerts, reminders, inferencing	PROforma Retrogram
Statistical and Probabilistic methods	Decision analysis methods, multiple regression, discriminate analysis, Bayesian belief networks	diagnostic support, therapy evaluation	Markov Models
Heuristics	Expert systems	Diagnostic and therapy support and evaluation, expert systems	Isabel DXPlain
Mathematical models	Discrete and continuous simulation models, optimization models	Modeling complex biological and/or physiological systems	ULAM*
Computational intelligence	Artificial neural networks (ANN), genetic algorithms, fuzzy logic	Diagnostic pathology, therapy evaluations, clinical guidelines	BASS** PAIRS***

Note: Many of the systems listed here use more than one methods.
**BASS: Biopsy Analysis Support System based on a modular neural network architecture.*
***PAIRS: Physician assistant Artificial Intelligence System.*
****ULAM: UNOS Liver Allocation Model, which uses discrete event simulation models to model the process by which transplant candidates and organs are listed.*

Statistical and probabilistic approaches to medical diagnosis dates back to early 1960s. Warner and his colleagues developed one of the first medical applications to use a decision support system based on Bayesian probabilities to describe how to change one's beliefs in light of new evidence, and to incorporate concepts like uncertainty and degrees of beliefs (Warner, 1961). Other relevant statistical methods have also been applied (e.g., multidimensional classification techniques) such as cluster analysis, where a case is identified to belong to a particular class (e.g., diagnosis) based on the values of certain parameters.

Unlike a probabilistic methodology, however, systems that employ heuristic modeling are based on a clinician's experience and thus present evaluations that are easier for other clinician's to interpret. Heuristics in this context attempts to capture elements of human expertise and cognition as a basis for modeling the reasoning process. Because such a methodology uses expert knowledge or heuristics in the form of rules, they are also referred to as *rule-based systems* or *expert systems*. MYCIN was the first instance of a medical application founded on the idea of creating production rules based on heuristic modeling and was designed to assist with the selection of appropriate antimicrobial therapies (Shortliffe & Buchanan, 1990). A recently developed CDSS, known as Isabel, can work with electronic health records, and encapsulates over 6,500 diagnoses (Ramnarayan, Maheshwari, Britto, Vyas, Wilson, & Thomas, 2003). Isabel uses heuristic rules to filter diagnoses based on age, gender, and other groupings. The DXplain system has been in use for over 20 years, and uses heuristic scoring rules to develop its diagnoses.

Likewise, a variety of mathematical modeling methods have been used in CDSS over the past several decades. The objective of the model is to describe a complex biological or physiological system often through a discrete or continuous simulation model. Optimization models are also used to address questions such as optimum dosage.

Last but not least, computational intelligence refers to techniques such as artificial neural networks (ANN), fuzzy logic, and genetic algorithms. ANN's enable systems to learn from previously documented examples of disease, treatment and outcomes. For instance, neural networks accomplish this through recognizing patterns in the input data and classifying them. Examples of this technology include automatic interpretation of cardiac images and pulmonary gas exchange parameters (Peleg & Tu, 2006).

Completing the framework for understanding CDSS are a number of issues, as shown in Figure 2. The following sections provide a brief treatment of these issues.

CURRENT ISSUES

CDSS Integration and Adoption

Despite the practical and theoretical advances in CDS, widespread use is still not taking place in the U.S. There is, however, substantial evidence that CDS can provide enhanced safety, quality, adherence to clinical guideline-based care, improved patient monitoring and decrease in medication-related errors (Chaudhary, Wang, Wu, Maglione, Mojica, & Roth, 2006). Why then, has CDS not been more widely adopted in the healthcare community? Researchers such as Peleg and Tu (2006), who have investigated the success factors for CDS, emphasize that: CDS should be computer-based; CDS intervention should occur in a timely manner; and the costs and clinical effects should be consistently and quantitatively evaluated. Wetter (2002) further suggests that CDSSs must be flexible and respond to users needs, and clinicians should be able to update the knowledge base.

Integration of CDSSs into the healthcare IT environment has been slowed by the lack of computerized clinical data sources, such as Electronic Health Records (EHR). The healthcare industry has only recently begun utilizing EHRs as the standard for maintaining detailed health information.

CDSS Development

The increasing confidence in CDSSs to improve the quality and safety of healthcare delivery has encouraged considerable development in the technology in recent years. At present, development is focused on better understanding the medical reasoning process, developing knowledge representation standards, improving knowledge acquisition capability, machine learning, ontologies and clinical vocabulary standardization, among others. Developing robust CDSSs will require significant investments in research and product development activities, and healthcare IT vendors will need to respond with innovative strategies for product development, research, and evaluation methods.

CDSS Evaluation

There are a variety of ways to define evaluation. In its most basic form, CDSS evaluation is concerned with assessing the effectiveness of a decision support tool. Does the system function as required? Does it improve clinician performance and patient outcomes? How these questions should be answered is matter of ongoing debate. Two relevant research methodologies include: the *positivist*, natural science model that mandates quantitative evaluation of CDSSs, and the *interpretivist* approach, primary approach used in social sciences that attempts to understand and interpret the social and cultural context in which people and technology coexist. Evaluation frameworks that adopt both of these methodologies have been proposed, and may prove to be the most effective method for CDS evaluation.

The suggested benefits of CDS include improved patient safety, quality of care, and improved efficiency of healthcare delivery (Sintchenko, Gilbert, Coiera, & Dwyer, 2002). The nature of CDSS design, experimental as it is, complicates the matter of evaluation. Different software systems, different users, environments, vocabularies, and methods of representing knowledge and clinical guidelines contribute to the problem (Osheroff, Teich, Blackford, Steen, Wright, & Detmer, 2006). The assumptions underlying CDSS evaluations have to be considered carefully for accurate interpretation of the results. For example, in a 2003 study of 55 CDSS evaluations, it was found that less than 25% of the evaluations involved a randomized controlled trial, where users' before and after perceptions were emphasized to a greater degree (Sintchenko & Coiera, 2003).

Another marker that is essential to effectively evaluating CDSSs is patient outcome. Few studies to date have shown any significant clinical impact on patient outcome (Garg, Adhikari, McDonald, Rosas-Arellano, Devereaux, & Beyene, 2005). One could assume that once CDS is used on a large scale, the reduction of medical errors and improved safety and quality will indeed impact clinical outcomes. Nevertheless, until such benchmarks are developed through targeted research that quantitatively measures improved patient outcomes, evaluation of CDSSs will be lacking.

Legal and Ethical Issues

The use of CDSSs in practice has not yet been undertaken on a large scale. Because of this, the U.S. courts have not specifically considered the liability issues arising from their use (Rustad, 2004). This is certain to change as CDS becomes more widely adopted. Currently, of the more than ten thousand categories of medical devices in use in the U.S., more than half of them functionally rely on embedded software, and for this reason, consideration of the legal implications of the use of software is important (Gage & McCormick, 2004).

U.S. Tort law, which differentiates liability issues into two camps—negligence and strict liability—represent the most generally accepted legal approaches available to the plaintiff (Miller, Schaffner, & Meisel, 1985). In medicine, negligence refers to treatment that does not follow the established standards of the medical community. Strict liability, which is the basis for the canon of product liability in the U.S., only acknowledges whether or not the defendant's action caused harm, and does not consider the intentionality of the act.

By which standard will the individuals, manufacturers, and organizations involved in healthcare delivery be held? Who will assume responsibility for developing systems that harm patients, and how will the individuals and organizations take steps to ensure their systems are safe? These are some of the legal challenges facing CDS.

In many ways, the ethical considerations for CDS are similar to other issues raised with respect to the use of technology in medicine. As consumers of healthcare we are concerned about our privacy as patient health records are digitized and integrated with billing, insurance and other business functions. As with other types of technological medical interventions, CDSS poses a risk to the misuse of health information. The capacity for CDSSs to contain vast quantities of knowledge, health-related statistics, prognoses and outcomes, greatly complicates the scenario. Who will have access to this information, and how will it be used? Can we assume that everyone in the healthcare field has our best interests at heart? How will genetic data and related health information impact employability or insurability? The risk of bias and discrimination based on our health information is a matter of serious concern. CDSS developers and users must be careful to ensure that the rapid pace of development and implementation of CDSSs does not overshadow the potential to do harm. Goodman (2007) suggests that CDSS users need to be educated extensively on its use, and points out the potential for emergence of ethical issues when systems are not used for their intended purpose.

FUTURE TRENDS

To realize the potential of CDS, there are several key research and development areas worthy of attention:

1. **Development of standardized best practice implementation strategies for CDSS:** The variety of types of CDSSs in use along with a lack of standards for features, vocabularies, and methods of representing knowledge present challenges for successful implementation of CDSSs.

2. **Development of standardized CDSS interventions and integration strategies for focused, high-impact clinical conditions:** Focusing our efforts initially on high-impact health conditions such as diabetes, CHF, asthma, and so on will place CDSS in a position to have a substantive economic impact in terms of improvement of chronic health conditions, disease maintenance and the associated cost-savings.
3. **CDSS workflow studies:** Modern CDSSs have tended to move away from their original goal of producing stand alone decision tools. The attention now is on building highly focused, interactive systems that are placed contextually within the complex and demanding workflow of today's clinician, as well as the healthcare information system as a whole, so as to provide more opportunities to prevent errors (Garg et al., 2005).
4. **Central clinical knowledge repository research:** Investment pertaining to the development of a medical community-based, noncommercial central knowledge repository would facilitate the organization, development and standardization of medical knowledge that could be accessed by a variety of systems using a variety of technologies to provide CDS.
5. **Development of standards for the practical representation of clinical knowledge and data:** Over the past decade, ontologies have been used to symbolically represent and reason about clinical knowledge, thus playing a vital role in the establishment of clinical practice guidelines. When represented in a formal language such as frames or description logic, relationships can be identified in existing data and inferences made (Peleg & Tu, 2006).
6. **Development of best practices for documenting and assessing the quality of CDSS content:** Physician performance, patient outcomes and system use ultimately reflect the quality of CDSS content. Research specifically designed to develop best practices for documentation and assessment of content quality, as well as the impact of that quality, is needed to ensure that CDSS is relevant, safe, and contributing to the improvement of healthcare.

CONCLUSION

In sum, there is significant interest in the use of CDSSs to support clinical decision tasks. CDSSs have been shown to reduce medical errors, improve clinician performance, and increase the quality of care. CDSSs employ a variety of methods and formalisms, each of which contributes in unique ways. There are a number of issues surrounding the use of CDSSs, including development, implementation and adoption, ethical and legal concerns, and evaluation issues, which need further attention from researchers and practitioners.

REFERENCES

Berner, E. S., & La Lande, T. J. (2007). Overview of clinical decision support systems In E. S. Berner (Ed.), *Clinical Decision Support Systems: Theory and Practice* (2nd ed., pp. 3–22). Springer.

Chaudhary, B., Wang, J., Wu, S., Maglione, M., Mojica, W., & Roth, E. (2006). Impact of information technology on quality, efficiency, and costs of medical care. *Annals of Internal Medicine, 144,* 742–752.

Degoulet, P., & Fieschi, M. (1997). *Introduction to Clinical Informatics: Theory and Practice* (Second ed.). Springer.

Degoulet, P., Jean, F. C., Engelmann, U., Meinzer, H. P., Baud, R., Jagermann, C., et al. (1995). Medical application development and the HELIOS software engineering environment. In H. U. Prokosch, & J. Dudeck (Eds.), *Hospital Information Systems: Design and Development Characteristics, Impact and Future Architecture* (pp. 3–15). Amsterdam: Elsevier.

Gage, D., & McCormick, J. (2004). *Can software kill?* Retrieved February 12, 2008, from http://www.eweek.com/article2/0,1759,1544225,00.asp

Garg, A. X., Adhikari, N. K., McDonald, H., Rosas-Arellano, M. P., Devereaux, P. J., & Beyene, J. (2005). Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: A systematic review. *Journal of the American Medical Association, 293,* 1223–1238.

Goodman, K. W. (2007). Ethical and legal issues in decision support. In E. S. Berner (Ed.), *Clinical Decision Support Systems: Theory and Practice* (pp. 126–139). Springer.

Greenes, R. A. (2006). A brief history of clinical decision support: Technical, social, cultural, economic, and governmental perspectives. In R. A. Greenes (Ed.), *Clinical Decision Support: The Road Ahead* (pp. 31–77). Academic Press.

IOM. (1999). *To err is human: Building a safer health system*. Institute of Medicine.

IOM. (2001). *Crossing the quality chasm: A new health system for the 21st century*. Institute of Medicine.

Ledley, R. S., & Lusted, L. B. (1959). Reasoning foundations of medical diagnosis; symbolic logic, probability and value theory aid our understanding of how physicians reason. *Science*, *130*(3366), 9–21.

Miller, R. A., Schaffner, K. F., & Meisel, A. (1985). Ethical and legal Issues related to the use of computer programs in clinical medicine. *Annals of Internal Medicine, 102,* 529–536.

Musen, M. A., Shahar, Y., & Shortliffe, E. H. (Eds.). (2001). *Medical informatics: computer applications in healthcare and biomedicine.* New York: Springer.

Osheroff, J. A., Teich, J. M., Blackford, M., Steen, E. B., Wright, A., & Detmer, D. E. (2006). A roadmap for national action on clinical decision support. *Journal of the American Medical Informatics Association, 14*, 141–145.

Peleg, M., & Tu, S. (2006). Decision support, knowledge representation and management in medicine. In *IMIA Yearbook of Medical Informatics 2006* (pp. 72-80).

Ramnarayan, P., Maheshwari, P., Britto, J. F., Vyas, H., Wilson, S., Thomas, N. J., et al. (2003). Evaluating the performance of a pediatric differential diagnostic tool (ISABEL) in critical care: A USA-UK multicentre study. *Critical Care Medicine, 31*(12).

Rustad, M. L. (2004). Punitive damages in cyberspace: Where in the world is the consumer? *Chapman Law Review, 7,* 39.

Shortliffe, E. H., & Buchanan, B. G. (1990). A model of inexact reasoning in medicine. In *Readings in uncertain reasoning* (pp. 259–275). San Francisco, CA: Morgan Kaufmann Publishers Inc.

Sintchenko, V., & Coiera, E. W. (2003). Which clinical decisions benefit from automation? A task complexity approach. *International Journal of Medical Informatics, 70*(2–3), 309–316.

Sintchenko, V., Gilbert, G. L., Coiera, E., & Dwyer, D. (2002). Treat or test first? Decision analysis of empirical antiviral treatment of influenza virus infection versus treatment based on rapid test results. *Journal of Clinical Virology, 25*(1), 15–21.

Warner, H. R. (1961). Mathematical approach to medical diagnosis. *Journal of the American Medical Association*(177), 75–81.

Wetter, T. (2002). Lessons learnt from bringing knowledge-based decision support into routine use. *Artificial Intelligence in Medicine, 24*(3), 195–203.

KEY TERMS

Active CDSS: Active CDSSs make decisions and interact without input or request by the user.

Artificial Neural Networks: A network of simple processing elements which can exhibit complex global behavior, determined by the connections between the processing elements and element parameters.

Fuzzy Logic: Based on fuzzy set theory, deals with complex systems where reasoning is approximate, due to complexity or incomplete data.

Genetic Algorithms: A genetic algorithm is a search method used in computational intelligence to find true or approximate solutions to optimization and search problems.

Heuristic: An algorithmic technique designed to solve a problem that ignores whether the solution can be proven to be correct.

Inference Engine: The component of a CDSS program that formulates inferences from a knowledge repository.

Passive CDSS: Requires the user to initiate a process by providing a request to the system.

C

Current Source Design for Electrical Bioimpedance Spectroscopy

Fernando Seoane
University College of Borås, Sweden

Ramon Bragós
Technical University of Catalonia, Barcelona, Spain

Kaj Lindecrantz
University College of Borås, Sweden

Pere J. Riu
Technical University of Catalonia, Barcelona, Spain

INTRODUCTION

The passive electrical properties of biological tissue have been studied since the 1920s, and with time, the use of Electrical Bioimpedance (EBI) in medicine has successfully spread (Schwan, 1999). Since the electrical properties of tissue are frequency-dependent (Schwan, 1957), observations of the bioimpedance spectrum have created the discipline of Electrical Impedance Spectroscopy (EIS), a discipline that has experienced a development closely related to the progress of electronic instrumentation and the dissemination of EBI technology through medicine.

Historically, the main developments in EIS related to electronic instrumentation have been two: first, the progressive shift from "real studies," where only resistance is measured, to "complex studies," where the reactance is also measured. Second, the increasing upper limit of the measurement frequency makes it possible to perform studies in the whole β-dispersion range (Schwan, 1957).

Basically, an EBI measurement system obtains the relationship between voltage and current in an object, obtaining the impedance or the admittance according to Equations 1 and 2. In EBI, most of the systems measure the impedance of the tissue, therefore injecting a known current and measuring the corresponding voltage drop in the biological sample. See Figure 1(A).

$$\boldsymbol{Z} = \frac{\boldsymbol{V}}{\boldsymbol{I}} \tag{1}$$

$$\boldsymbol{Y} = \frac{\boldsymbol{I}}{\boldsymbol{V}} \tag{2}$$

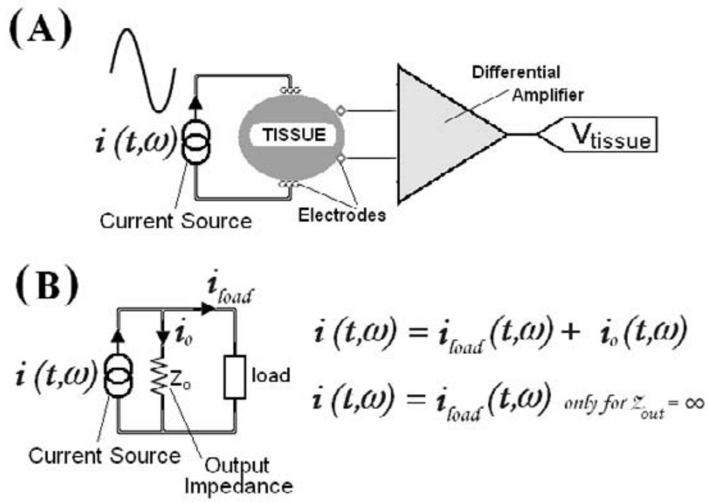
In an EBI measurement system, the current source plays a very important role, and its features are critical for the overall performance of the measurement system—especially for the frequency range of operation and the load dynamic range that the system is able to measure. The use of current driving instead of voltage driving is the most extended approach (Morucci & Rigaud, 1996), which applies an intrinsically safe current-limiting mechanism and reduces the possible nonlinearities. Given that the signal generator usually provides a voltage at its output, the current driver often is a Voltage Controlled Current Source (VCCS).

BACKGROUND

The functional purpose of a current source is to generate an electric current signal with a specific magnitude. Therefore, the output current is the most important characteristic of a current source, and parameters related to the output current define the performance of the source.

An ideal current source will provide exactly the same current to any load independently of its value at any frequency. To fulfil this aim, the output impedance Z_{out} must be very large, ideally ∞, at all frequencies. See Figure 1(B).

Figure 1. (A) An ideal representation of a four-terminal voltage over current EBI measurement system; (B) representation of an ideal current source



In practice, the frequency range is limited to an operational frequency range at which the value of the output impedance of the source is very large in comparison with the value of the load. In bioimpedance applications, a commonly accepted value is at least 100 kΩ.

As a safety measure in biomedical applications, when current is injected into the body, the accumulation of electrical charges is avoided as much as possible. Therefore, the DC component of the stimulating current should be zero.

Over the years, there have been two main approaches in current source design: voltage-based structures and current-based structures. In voltage-based structures, what is responsible for the generation of the output current is the voltage in one or more nodes of the active circuit (e.g., the virtual ground in the case of Op-Amps). In a current-based structure, the output current is generated by an active device with intrinsic current-mode operation (e.g., Transconductance Amplifiers and Current Conveyors).

Voltage-Based Structures

Howland-based circuits and Load-in-the-Loop structures are the most common approaches to implement VCCS, and both are based on a single Op-Amp circuit. Another family of VCCS is based on a differential amplifier with unity gain and positive feedback.

Load-in-the-Loop

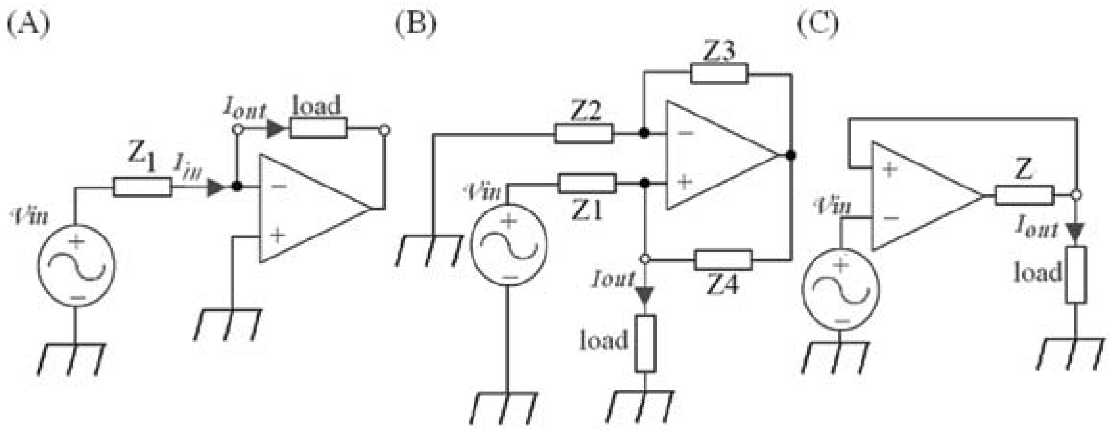
This current source design is one of the approaches first used to implement current sources with floating load (Sheingold, 1966). In Figure 2(A) it is possible to observe the principle of operation for the output current generation and the importance of the Op-Amp circuit. Note that as long as the Op-Amp provides a virtual ground and infinite input impedance, the output current is independent of the value of the load (i.e., the output impedance is infinite). Since the characteristics of an Op-Amp circuit are frequency-dependent, the output impedance of the current source exhibits a similar dependence. Equation 3 shows the analytical expression for the output impedance, Z_{out}, of the current source circuit in Figure 2(A), considering the frequency dependence of the differential gain (Seoane et al., 2007). Z_0 and Z_{in} are the output and input Op-Amp impedances, and Z_i is the impedance which determines the transconductance of the VCCS, Z_1 in Figure 2(A).

$$\boldsymbol{Z_{out}} = \left(\boldsymbol{Z_o} + \left(\boldsymbol{Z_{in}} \mathbin{//} \boldsymbol{Z_1}\right)\left(\boldsymbol{A_d(s)}+\mathbf{1}\right)\left(\mathbf{1} - \frac{\mathbf{1}}{\mathbf{2CMRR(s)}}\right)\right) \tag{3}$$

Howland Source

This circuit topology (Sheingold, 1964) is probably the structure most used as a current source in EBI. Figure

Figure 2. (A) A Voltage Controlled Current Source implementation with Load-in-the-Loop. (B) A Voltage Controlled Current Source implemented with the original Howland current circuit. (C) VCCS based on a differential amplifier with positive feedback



2(B) shows the original circuit for the Howland source. When the impedance bridge is balanced (i.e., Z_1 x Z_3 = Z_4 x Z_2), the output current is proportional to V_{in} and only dependent on the value of Z_2 (4). Regarding the output impedance, as long as the bridge is balanced the output impedance of the current source is infinite; see Equation 5.

$$I_{out} = \frac{V_{in}}{Z_2} \quad (4) \qquad Z_{out} = \frac{Z_1 \times Z_2 \times Z_4}{Z_2 \times Z_4 - Z_1 \times Z_3} \quad (5)$$

The expression in Equation 6 is the output impedance of the Howland source, considering the frequency dependence of the differential gain of the Op-Amp circuit.

$$Z_{out}(s) \left(1 \quad \frac{A_d(s)}{1+\frac{Z_3}{Z_1}} \right) \left(Z_1 \| Z_3 \right) \quad (6)$$

Even using perfectly matched resistors, there is a degradation of Z_{out} due to the Op-Amp frequency response. In addition, resistors' tolerance imposes a finite Z_{out} even at low frequencies. Note that, for the sake of simplification of the final expression, it is assumed that $Z_2 = Z_1$ and $Z_4 = Z_3$, which also balance the bridge.

The Howland circuit has been studied for many years and enhancements on this topology have been implemented and tested (Bertemes-Filho, Brown, & Wilson, 2000).

Structures Based on a Differential Amplifier

Several VCCS structures can be reduced to a differential amplifier with unity gain and positive feedback, as depicted in Figure 2(C). As shown in Equation 7, in the case where the amplifier differential gain A_d is equal to 1, the output current I_{out} is load-independent, (i.e., the output impedance of the current source is infinite).

$$I_L = \frac{(A_d - 1) \times V_L - A_d \times V_{in}}{Z} \Rightarrow I_L \Big|_{A_d = 1} = -\frac{V_{in}}{Z} \quad (7)$$

The differential amplifier admits several implementations, monolithic or with discrete Op-Amps, and typically with classical structures having one, two, or three devices. The output impedance of the resulting VCCS is not ideal, and depends on the accuracy of the condition $A_d = 1$, as shown in Equation 8.

$$Z_{out}(s) = \frac{Z}{1 - A_d(s)} = \frac{Z}{\varepsilon} \frac{1+\frac{s}{\omega_o}}{1+\frac{s}{\omega_0 \varepsilon}} \quad (8)$$

At very low frequency, its value is $Z_0 = Z/\varepsilon$, with ε being the relative error in A_d. The first pole of Z_0 depends on the first pole of A_d and again on ε. Both this error and its frequency behaviour will depend on the structure used to build the differential amplifier, on the tolerance of devices, and on the frequency response and the common mode rejection ratio (CMRR) of the Op-Amps.

Current-Based Structures

Current sources based on current-mode structures (Toumazou, Lidgey, & Haigh, 1989) involve a transconductor at device level (e.g., transistor) or at circuit level. The resulting VCCS circuits normally use less passive components and are simpler than those based on Op-Amps, making them more suitable for integrated implementations. Their frequency bandwidth is higher than that obtained with Op-Amps based on the same technology. Most circuits based on current-mode devices are open-loop structures, which limits their accuracy. The core of circuit-level current-mode structures is usually a current-conveyor (CCII) or an operational transconductance amplifier (OTA). Both could be directly used to build VCCS circuits.

Current Conveyors

The current conveyor (CCII) (Sedra, Roberts, & Gohh, 1990) is a current-mode building block whose acronym stands for second generation current conveyor. CCII is a three-port device with two inputs (X, Y), and an output (Z). Figure 3(A) displays its symbol, and Figure 3(B), the simplified diagram of the usual implementation. The Y input is a high-impedance node whose voltage is copied to X, the low-impedance node. The input current at X node is copied to the Z high-impedance output node. To implement a VCCS with a CCII, a voltage generator V_i is connected to the Y node, and a resistor R between ground and the X node. Thus, node Z generates the current replicated at node X, providing a large output impedance. Note that, since $i_x = V_i/R = i_{out}$, R is the VCCS transconductance. The main limitation on its use is the lack of commercial availability (e.g., Analog Devices AD844, TI-Burr-Brown OPA660). There are several EBI and EIT instruments that include CCIIs in their structure (Casas, Rosell, Bragos, Lozano, & Riu, 1996; Yerworth, Bayford, Cusick, Conway, & Holder, 2002).

Transconductance Amplifiers

In essence, an operational transconductance amplifier (OTA) is an amplifier that generates an output current linearly proportional to its voltage differential inputs, as expressed in Equation 9. The gain of an OTA is not just gain, but transconductance, denoted by g_m and adjustable by the value of input currents I_{abc} and I_{bias}. See Figure 3(C).

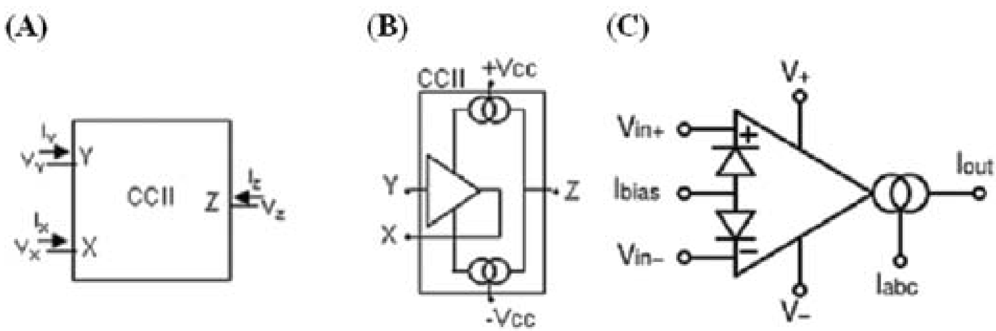
$$I_{out} = (v_{in+} - V_{in-}) \times g_m \qquad (9)$$

Since the OTA is a current output device, its output impedance should be very large. Therefore, considering its ideal features of output current controlled by voltage with a linear relationship and high output impedance, the OTA circuit is the perfect VCCS.

The Effect of Parasitic Capacitances

Parasitic capacitances can be found associated with many elements of an impedance measurement system.

Figure 3. (A) Current conveyor: symbol; (B) current conveyor simplified schematic; (C) OTA, schematic symbol for an operational transconductance amplifier



They may be associated with the output of the current source, C_o, the stimulating leads, C_m, and the sensing leads, C_{in}, as well as between the electrodes, C_{ie}, the system's ground and earth, C_{is} and even the patient and earth, C_{bg}.

All these parasitic capacitances create pathways for the current to leak away from the tissue for measurement (e.g., the patient). The origin of such capacitances and their effects have been studied in detail by several authors (eg., Scharfetter, Hartinger, Hinghofer-Szalkay, & Hutten, 1998).

The main effect of parasitic impedances associated with the current source is to reduce the output impedance of the source with frequency. This effect may be negligible in the frequency range of operation of the measurement system, but otherwise the parasitic capacitances set the frequency limit of operation of the impedance meter, especially in those without a reference current measurement. The output impedance can be affected severely by parasitic capacitances associated with the output, C_o (Seoane, Bragos, & Lindecrantz, 2007), but also by parasitic capacitances intrinsic to the current source circuitry (ibid.). In practice, circuits without active capacitance compensation exhibit a certain output capacitance, including CCII and OTA, typically of 3–5 pf. This fact limits the output impedance to 50 kΩ at 100 kHz.

CHALLENGES AND DESIGN TRENDS FOR CURRENT SOURCES IN EIS SYSTEMS

The selection of a specific approach for the design of the current source reduces, in most of the cases, the versatility of the measurement instrument. Therefore, current source design usually is application-specific. Currently, electronic instrumentation for wideband multifrequency measurements is an important research area within electrical bioimpedance spectroscopy, pursuing the goal to widen the frequency band of operation of the EIS measurement systems.

Multifrequency Measurements

Biological tissue, due to its structure and the electrical properties of its constituents, presents an electrical impedance that varies with frequency (Schwan, 1957). Therefore, the impedance at a certain frequency is often different from that at another frequency, and for the same reason the impedance spectrum of a tissue or subject may provide information regarding the status of the tissue or its composition (e.g., body composition and skin cancer screening).

Currently, there are two methods to obtain the impedance spectrum of a tissue: using true multifrequency systems or sweeping frequency systems. The latter uses an excitation signal containing just one tone, sometimes two or even three, at a specific frequency. The spectrum impedance is obtained after several excitations by sweeping the frequency of the applied tone to cover the frequency range of measurement. In contrast, multifrequency systems use an excitation signal containing several tones, often a multisine (Bragos, Blanco-Enrich, Casas, & Rosell, 2001), obtaining the impedance spectrum of the sample after only a single excitation.

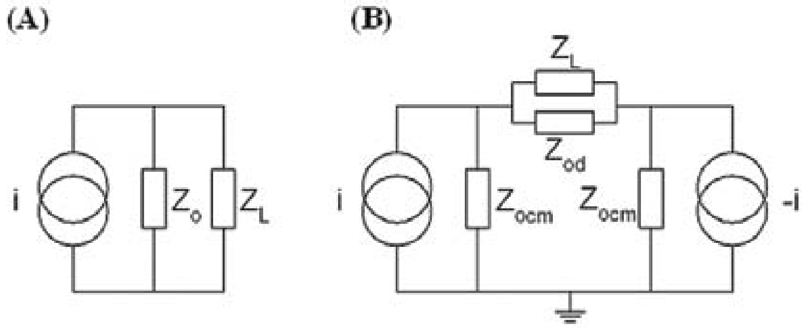
These two methods impose different requirements on the current source. Multifrequency systems need a current source able to provide a large output impedance simultaneously in the complete frequency range of measurement, while current sources for sweeping systems only need to provide a large output impedance at the frequency of the tone or tones contained in the excitation signal.

Enhancing Design Approaches

There are many factors that influence the performance of an impedance measurement system, and each type of current source exhibits an intrinsic robustness to different sources of errors. For instance, in previous paragraphs we have introduced the effect of parasitic capacitances that shunt the stimulating current away from the measurement tissue or patient. Just by selecting one type of current source, the robustness against certain parasitic effect can be improved (e.g., load-on-the-loop structures are more robust against leads' parasitic capacitance than grounded-load structures, like Howland). Such improvement is often slight, although in some cases, it is enough to obtain a suitable current source for a specific application.

In most cases, circuit design approaches for current sources are used to specifically minimize or even eliminate sources of error in EBI systems, such as parasitic capacitances.

Figure 4. (A) Floating current source; (B) implementation based on two equal sources referred to ground



Symmetrical Current Sources

The use of symmetrical current sources allows us to minimize the common mode voltage at the load, thereby reducing the errors, due to limited CMRR of the voltage measuring differential amplifier. The previously described current sources are referred to ground, except the load-in-the-loop source, which is floating but not symmetrical. Symmetrical current sources are presently implemented by connecting the load between the outputs of two complementary current sources as in Figure 4. The current at a given moment is injected by one of the sources and drained by the other one. The unavoidable impairment between both sources creates a differential current which finds a path to ground through the common mode output impedance Z_{Ocm} of the current sources. This fact would generate again a large common mode voltage at the load, unless a common mode feedback (CMFB) circuit is used (Casas et al., 1996, Goovaerts, Faes, Raaijmakers, & Heethaar, 1999).

Negative Impedance Converters

There are several different circuit topologies to implement: Negative Impedance Converters (NICs). Figure 5(A) shows a well-known NIC topology (Sedra & Brackett, 1979), widely used in EIS and EIT (Lee, Cho, Oh, & Woo, 2006; Ross, Saulnier, Newell, & Isaacson, 2003). Such topology is able to synthesize an inductance between the points *a* and *b* with the value given by Equation 10.

$$L = \frac{R_1 \times R_3 \times Z_4 \times C_1}{Z_2} \qquad (10)$$

Therefore, by connecting the NIC circuit in parallel to the output of the current source as in Figure 5(B), it is possible to obtain an equivalent output impedance that is only resistive, eliminating any capacitance associated with the output Z_{eq}, independently of its origin. Tuning the discrete components of the NIC to make *L* comply with Equation 11, the capacitive part of *Zeq* in Figure 5(B) can be cancelled at any specific frequency, ω_S. This yields at ω_S an output impedance only real and significantly large, in the order of Ω (Ross et al., 2003).

$$L\Big|_{IM\{Z_{eq}\}=0} = \frac{1}{\omega^2 \left(C_o + C_{par}\right)} \qquad (11)$$

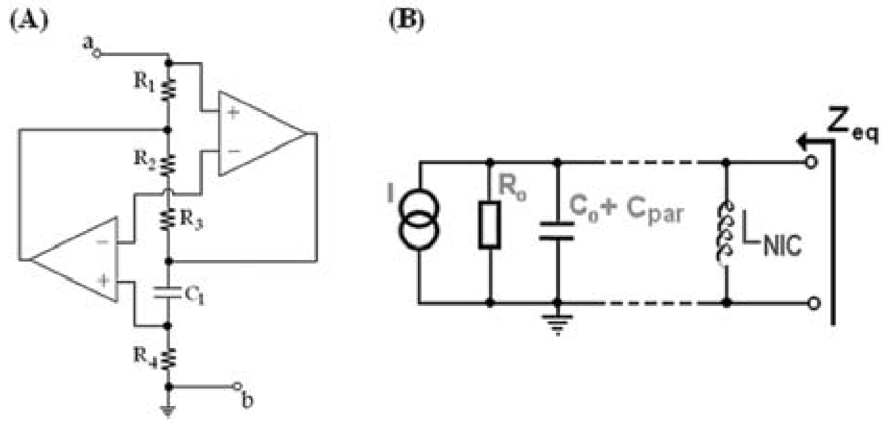
Despite the possibility to obtain a current source with virtually no reactive component, the use of NIC**s** is limited for several reasons:

- An NIC increases the instability of the system.
- NIC circuits need to be trimmed for each frequency before each measurement, limiting the use of the VCCS to only sweeping systems.
- The trimming process introduces unpractical time delays between measurements, and when the measurement system incorporates an automatic trimming block, such a block may be as complex as the rest of the measurement system.

Single Op-Amp Basis

Some of the sophisticated alternatives in the VCCS design appeared to overcome the drawbacks of single Op-Amp circuits implemented with the available Op-Amps at that time. Modern voltage and current-feedback

Figure 5. (A) Circuit schematic for a generalized impedance converter for inductances. (B) Equivalent circuit for the connection of the NIC circuit to the output of the current source.



amplifiers, with gain-bandwidth products of hundreds of MHz, allow us to retrieve these structures with good results for not highly demanding applications.

Current Conveyor + Load-in-the-Loop

Combining basic structures in cascade, where the CCII feeds current to a secondary current source, it is possible to obtain a VCCS circuit that is simple and has extraordinary performance, exhibiting large output impedance values at frequencies higher than 1 MHz (Seoane et al., 2007). The simplicity of this approach allows an easy microelectronic implementation avoiding any parasitic capacitance associated with internal connections and discrete components of the VCCS. The drawback of this design is that it does not allow the implementation of a symmetric current source; thus, it requires an external CMFB circuit to minimize the common voltage at the load.

Differential Difference Amplifier

A modern approach to the implementation of the VCCS based on a differential amplifier with unity gain is that which uses a Differential Difference Amplifier (DDA) to implement the differential amplifier. This structure allows high bandwidth and the unity gain is not compromised by resistors matching. (Ramos-Castro & Bragos-Bardia, 2004) describe an isolated front-end for cardiac applications whose current source is built with the AD830. (Riu, Anton, & Bragos, 2006) describe a symmetrical current source based on differential amplifiers with unity gain built around AD8130 with intrinsically high pass response, which ensures lower transconductance at low frequencies to improve safety.

FUTURE DEVELOPMENT TRENDS

Modern mixed-mode ICs (e.g., A/D and D/A converters, DDS, programmable gain amplifiers) usually have differential inputs and outputs, including current outputs in some cases. Differential floating architectures can then be more easily implemented at whole system level. Integrated implementations (Terzopoulos, Hayatleh, Hart, Lidgey, & McLeod, 2005) will help to reduce parasitic capacitances, but not the intrinsic output capacitance related to devices' loop-gain falling at high frequency. Only active capacitance compensation could allow a further improvement of output impedance. Proposed structures are, however, useful at only a single or a discrete set of frequencies, when using a parallel array of NIC circuits (Lee et al., 2006).

CONCLUSION

The incessant developments in electronic technology, especially in circuit integration and bandwidth of operation, together with the increasing numbers of applications making use of electrical bioimpedance spectroscopy, guarantee continuous improvement in the performance of electronic instrumentation for electrical bioimpedance measurements, including the current source. Classical structures can be retrieved with enhanced performance, and new building blocks allow wide operational bandwidths and large output resistances. Parasitic output capacitance is still a bottleneck, unless active capacitance compensation is used, but the reduced bandwidth of this technique limits its application in multifrequency EIS.

REFERENCES

Bertemes-Filho, P., Brown, B. H., & Wilson, A. J. (2000). A comparison of modified Howland circuits as current generators with current mirror type circuits. *Physiological Measurement, 21*(1), 1–6.

Bragos, R., Blanco-Enrich, R., Casas, O., & Rosell, J. (2001). *Characterisation of dynamic biologic systems using multisine based impedance spectroscopy*. Budapest, Hungary.

Casas, O., Rosell, J., Bragos, R., Lozano, A., & Riu, P. J. (1996). A parallel broadband real-time system for electrical impedance tomography. *Physiological Measurement, 17*, 1–6.

Goovaerts, H. G., Faes, T. J. C., Raaijmakers, E., & Heethaar, R. M. (1999). Some design concepts for electrical impedance measurement. *Annals of the New York Academy of Sciences, 873*(1), 388–395.

Lee, K. H., Cho, S. P., Oh, T. I., & Woo, E. J. (2006). *Constant current source for a multi-frequency EIT system with 10Hz to 500kHz operating frequency.* Paper presented at the IFMBE World Congress on Medical Physics and Biomedical Engineering 2006, Seoul, Korea.

Morucci, J. P., & Rigaud, B. (1996). Bioelectrical impedance techniques in medicine. Part III: Impedance imaging. Third section: Medical applications. *Crit Rev Biomed Eng, 24*(4–6), 655–677.

Ramos-Castro, J., Bragos-Bardia, R., et al. (2004). *Multiparametric measurement system for detection of cardiac graft rejection*. Como, Italy.

Riu, P., Anton, D., & Bragos, R. (2006). *Wideband Current Source Structures for EIT.* Paper presented at the IFMBE World Congress on Medical Physics and Biomedical Engineering 2006, Seoul.

Ross, A. S., Saulnier, G. J., Newell, J. C., & Isaacson, D. (2003). Current source design for electrical impedance tomography. *Physiological Measurement, 24*(2), 509–516.

Scharfetter, H., Hartinger, P., Hinghofer-Szalkay, H., & Hutten, H. (1998). A model of artefacts produced by stray capacitance during whole body or segmental bioimpedance spectroscopy. *Physiological Measurement, 19*(2), 247–261.

Schwan, H. P. (1957). Electrical properties of tissue and cell suspensions. *Adv Biol Med Phys, 5*, 147–209.

Schwan, H. P. (1999). The practical success of impedance techniques from an historical perspective. *Ann N Y Acad Sci, 873*, 1–12.

Sedra, A. S., & Brackett, P. O. (1979). *Filter theory and design: Active and passive*. London: Matrix.

Sedra, A. S., Roberts, G. W., & Gohh, F. (1990). Current conveyor. History, progress and new results. In *IEE Proceedings, Part G: Electronic Circuits and Systems* (Vol. 137, No. 2, pp. 78–87).

Seoane, F., Bragos, R., & Lindecrantz, K. (2007). *Current source for wideband multifrequency electrical bioimpedance measurements*. Unpublished manuscript.

Sheingold, D. (1964). Impedance and admittance transformations using operational ampliifers. *The Lightning Empiricist, 12*(1), 4.

Sheingold, D. (Ed.). (1966). *Applications manual for computing amplifiers for modeling, measuring, manipulating and much else.* Boston: Philbrick Researchers Inc.

Terzopoulos, N., Hayatleh, K., Hart, B., Lidgey, F. J., & McLeod, C. (2005). A novel bipolar-drive circuit for medical applications. *Physiological Measurement, 26*(5), N21–N27.

Toumazou, C., Lidgey, F. J., & Haigh, D. G. (1989). *Analogue IC design: The current-mode approach*. London, UK: Peter Peregrinus - IEEE.

Yerworth, R. J., Bayford, R. H., Cusick, G., Conway, M., & Holder, D. S. (2002). Design and performance of the UCLH mark 1b 64 channel electrical impedance tomography (EIT) system, optimized for imaging brain function. *Physiological Measurement, 1*, 149–158.

KEY TERMS

Current Source: A two-terminal analog electronic building block which generates an electrical signal with constant current amplitude independently of the load connected at the terminals.

Differential Amplifier: An amplifier whose output depends on the difference between two inputs through a gain (differential gain), and residually on the mean voltage between both inputs.

Electrical Bioimpedance: The physical magnitude that indicates the total impediment that a biomaterial offers to the flow of free electrical charges and the orientation of bounded electrical charges towards an existing electrical field.

Negative Impedance Converter: Type of electric circuit that can generate any impedance between two points—capacitances or inductances.

Operational Amplifier: Differential amplifier with a high open-loop gain which allows the implementation of accurate circuits by using voltage feedback.

Parasitic Capacitance: A capacitance usually defined between a node and ground, due to wires, tracks, pads, and p-n junctions in the signal path.

Transconductance: A contraction of *transfer conductance*, the relation between the voltage at the input of an electric system and the current at the output. It is denoted by g_m and measured in Siemens units.

The Current State of Bionanotechnology

Shivani Choudhary
Illinois Institute of Technology, USA

Eliezer Geisler
Illinois Institute of Technology, USA

Nilmini Wickramasinghe
Illinois Institute of Technology, USA

INTRODUCTION

Bionanotechnology is a combination of three terms: "bios" meaning "life," "nano" (origin in Greek) meaning "dwarf," and "*technologia*" (origin in Greek—comprised of "*techne*" meaning "craft," and "*logia*" meaning "saying"), which is a broad term dealing with the use and knowledge of humanity's tools and crafts. Biomolecular Nanotechnology—or Bionanotechnology—is a term coined for the area of study where nanotechnology has applications in the field of biology and medical sciences. One can also say that "Bionanotechnology" is derived by the combination of two terms: "nanotechnology," and "biotechnology."

BACKGROUND

The principles of physics, as far as I can see, do not speak against the possibility of maneuvering things atom by atom. It is not an attempt to violate any laws; it is something, in principle, that can be done; but in practice, it has not been done because we are too big.

By saying the above lines, Richard Feynman, Nobel Prize winner in Physics, clearly expressed the importance of decreasing the size of a material to very small scale.

In mathematics, "nano" is defined as one billionth of a meter or 10^{-9} meter. However, the term "nanotechnology" has no one definition attached to itself. If you ask a number of scientists and scholars to define the meaning of nanotechnology, you are bound to receive a range of definitions for the term. However, a basic definition of nanotechnology can also be stated as "**engineering of** functional systems at the molecular scale."[1] The following two lines help in expressing the idea of how small a nanometer really is:

1 meter = 1 billion nanometers[2]
Width of Human Hair = 80,000 nanometers[3]

Nanotechnology is often referred to as a general-purpose technology. That is because in its advanced form it will have significant impact on almost all industries and all areas of society. It offers better built, longer lasting, cleaner, safer, and smarter products for the home, for communications, for medicine, for transportation, for agriculture, and for industry in general.

Imagine a medical device that travels through the human body to seek out and destroy small clusters of cancerous cells before they can spread. Or a box no larger than a sugar cube that contains the entire contents of the Library of Congress. Or materials much lighter than steel that possess ten times as much strength.

- U.S. National Science Foundation

Nanotechnology being one of the most promising and growing industry today, a large amount of work that is being carried out in this area. Bionanotechnology is currently in its prime with researches being carried out all over the globe in this field. However, in the past few years, there also have been some great discoveries and inventions in this area of study.

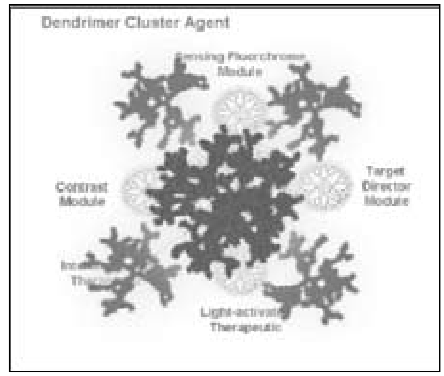
APPLICATIONS OF BIONANOTECHNOLOGY

Bionanotechnology deals with nanomaterials and their applications in life sciences. The applications[4,5] of bionanotechnology are vast. However, the bionanotechnology applications (Lin & Datar, 2006) can be broadly categorized into two main areas: therapeutics and diagnosis.

Nanotherapeutics

Nanotherapeutics is an application of bionanotechnology in which the diseases can be treated by the use of various techniques at molecular level. Nanotheraputics can be categorized into three main areas of study and development: *drug therapy, gene therapy, and immunotherapy*. However, the current research is being conducted mainly in the former two areas of nanotherapeutics. In *drug therapy*, the potential of water-insoluble drugs or unstable drugs can be greatly enhanced by reducing the size of the drug or by encapsulating the drug particle. In *gene therapy*, functional genes can be efficiently delivered to correct genetic disorders like hemophilia. This can be done by condensing the DNA into nanoparticles,[6] and hence controlling its composition for effective gene delivery.

Figure 1. Baker's dendrimer: Multifunctional cancer treatment (Courtesy: Nanotechnology takes on cancer)



Nanodiagnosis

Because of nanotechnology, diagnosis of a disease will become not only easy, but also highly effective. We will be able to detect a malfunction in even a single cell at the very moment it becomes defective. As a result, we will be able to diagnose a disease effectively in its prime stage which would thus result in an early treatment of the disease, like cancer (Grodzinski, Silver, & Molnar, 2006; Jain, 2005), hence increase the chances of survival.

The reason for early diagnosis due to the implementation of nanotechnology is that nanotechnology operates on the same scale as biology. Moreover, due to of the small size of nanoparticles, they will be able to gain access to areas of the body—such as the brain and individual cells—that have proved difficult to reach with current technologies.

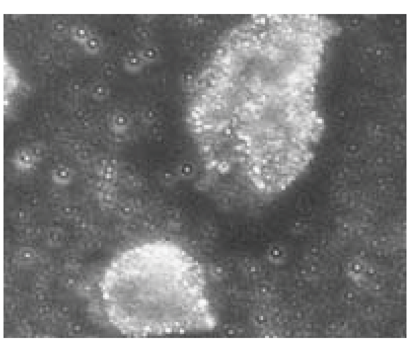
In the paragraphs to follow, there has been an attempt made in order to give an insight of what is currently being done in the field of Bionanotechnology.

- **Dendrimers:** (Svenson & Tomalia, 2005; Figure 1) It is a three-dimensional, branched nanoscale molecule resembling the structure of a "*tree*." The term "dendrimer" comes from a Greek word meaning "tree." It can act as a good *drug delivery agent* as well as a *time-release delivery agent*. It is normally used in therapeutic developments which are targeted to cancer cells (Baker, 2003). Researchers at Northwestern University are working on cancer detection, and stopping it before it starts using dendimers.
  Dendritic polymers are also used for the development of nanosensor/NEMS systems for noninvasive, continuous monitoring of astronauts for the biologic effects of space travel.
- **Patient-specific medicine (Baba, 2006):** Every individual has a different genetic structure which is responsible for things like color of the hair, eye's color, shape of nose, and so on. Hence, each human being has a distinct predisposition to a disease. Therefore, personalized medicine or patient-specific medicine simply means the *prescription of specific therapeutics best suited for an individual*. Xiaolian Gao, professor of chemistry, has developed a chemical process for building a device that could help doctors predict a patient's response to drugs or screen patients for thousands of genetic mutations and diseases, all with one simple lab test—on a DNA chip. The ultimate vision of this research is the development of patient-specific medicine.
- **Bionanosensors[7]:** Researchers in many universities and R&D laboratories are trying to develop sensors based on nanotechnology, which, when implanted under the skin of a human being, will be able to detect the level of glucose, hormones, or cholesterol. Nanosensors (Clark, Singer, Korns, & Smith, 2004; Jain, 2005; Kohli & Martin, 2005; Kvennefors & Persson, 2004) can play a vital role in the treatment of genetically based diseases like sickle cell anemia, due to their characteristics of

have high sensitivity and specificity. Moreover, the study of this kind of technology has demonstrated that the cancer cells can be easily identified. Recently, New Mexico Tech researchers have developed a nanosensor which can be used to analyze and identify individual components of single strands of DNA and RNA. This sensor is based on optical properties of nanoparticles.

- **Smart Drugs[8,9]:** Drugs when injected inside the human body will be able to predict the specific location where the human cells have been contaminated and then release a precisely targeted drug dose (Peppas). Belgian chemists have developed "self exploding" microcapsules which will be able to precisely release drugs inside the human body weeks or even months after injection.
- Biochemist Susan Hardin, and four University of Houston colleagues, are developing a new technology for direct molecular sensing that could be used to sequence an individual's entire genome—the gathering of all the genetic information contained in a person's DNA—in less than 24 hours, rather than 2–3 days. When fully developed, the technology could offer doctors a more rapid and more thorough way to determine who is at risk for certain genetic diseases, or which people might react adversely to a particular drug.
- **Gold nanoshells:** (Figure 2) They are nanoparticles made of silica (glass), which are completely coated with gold (Ha, Jeong, & Chung, 2005; Jain, 2005). Their diameter can vary from 100 nm to as large as several hundred nanometers. Researchers are trying to make use of gold nanoshells for the detection as well as treatment of cancerous cells.

Figure 2. Gold nanoparticle (Courtesy: When gold becomes a catalyst by staff writers)



- **Quantum dots[10]:** (Figure 3) "Quantum dots" are nanoparticles made out of semiconductors having unique electrical and optical properties (Chan, 2006; Jain 2005; Jiang, Papa, Fischer, Mardyani, & Chan, 2004; Weng & Ren, 2006). They emit distinct colored when exposed to light depending upon their size. The smaller the quantum dots, the brighter the color. They are specifically used for *imaging purposes* inside the human body. They demonstrate extremely high sensitivity. Imagine a single protein tracked in a living cell by the use of the optical properties of quantum dots.
- **Silver nanoparticles:** (Figure 4) As the term suggests, they are nanoparticles derived from silver atom. A study conducted by the University of Texas and Mexico University, found that silver nanoparticles can kill HIV-1, and is likely to kill virtually any other virus.

Figure 3. Electron micrographs reveal arrays, shown in false color, of spherical and cubic quantum dots made of lead selenide and lead telluride (Courtesy: A. Norman, S. Ahrenkiel, A. Hicks, J. Murphy/NREL)

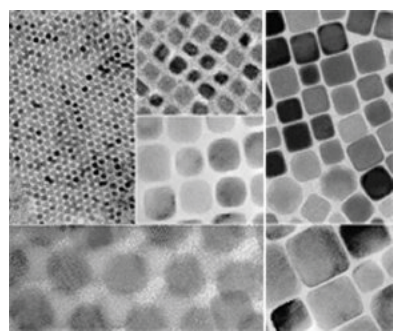
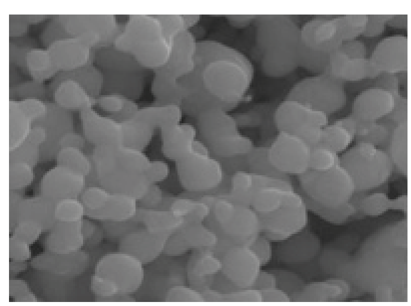


Figure 4. Silver nanoparticles



- **Regenerative medicines:** (Emerich, 2003; Rajangam, Behanna, Hui, Han, Hulvat, Lomasney, & Stupp, 2006; Yamato & Okano, 2006) Regenerative medicine helps natural healing processes to work faster, or uses special materials to regrow missing or damaged tissue. Scientists at Northwestern University are working towards the development of regenerative medicines. They have successfully *grown nerve cells* using an artificial three-dimensional network of nanofibers. They have also developed a liquid that forms a gel-like mass of nanofibres on contact with, water which could provide the most promising vehicle yet for the *regeneration of damaged spinal cords*. Many scientists working in the area of regenerative medicine are trying to develop *artificial scaffolds* that store or attract cells, and then control their growth and final identity.

FUTURE TRENDS OF BIONANOTECHNOLOGY

As we see, nanotechnology has a large number of applications in the field of biology and life sciences. Due to this, bionanotechnology is receiving huge amount of funds from not only the government, but also private lenders like the Venture Capitalists. In fact, bionanotechnology has become one of the top areas of funding by the Venture Capitalists.

The future trends of nanotechnology seem to be definitely increasing. Medical device applications, research, technological applications, and so on are being increasingly approved and funded. "*As of mid-2006, 130 nanotech-based drugs and delivery systems and 125 devices or diagnostic tests are in preclinical, clinical or commercial development.*"[11] According to the U.S. National Science Foundation (NSF) (2001), by 2015, about a half of all the pharmaceutical industry products, which is over $180 billion per annum, will be produced by nanotechnology.

Many market researches have been conducted on the future of bionanotechnology. Here we shall describe only a few.

Research conducted by *Lux Research Inc.*[12] (Figure 5 & 6) shows that R&D, conducted throughout the globe, in all the sectors of nanotechnology approximated around $9.6 billion in 2005. It also projected that share of nano-enabled drug delivery in the market will rise from $980 million in 2005 to approximately $8.6 billion by 2010. Nanotherapeutics approximated $28 million in 2005, and is bound to grow to $310 million by 2010, while the nano-enabled diagnostics which were $56 million in 2005 will grow to about $1 billion by 2010.

Figure 5. Results of research conducted by Lux Research Inc.

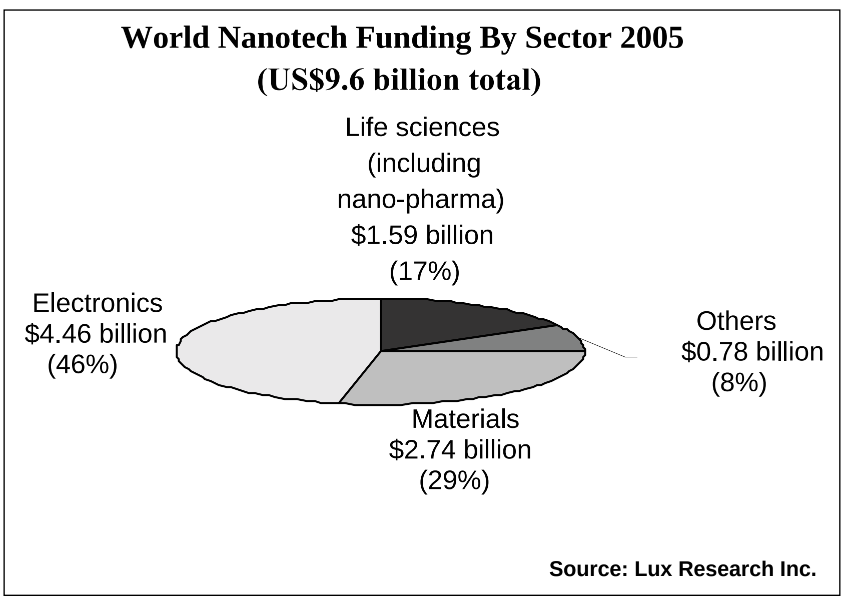
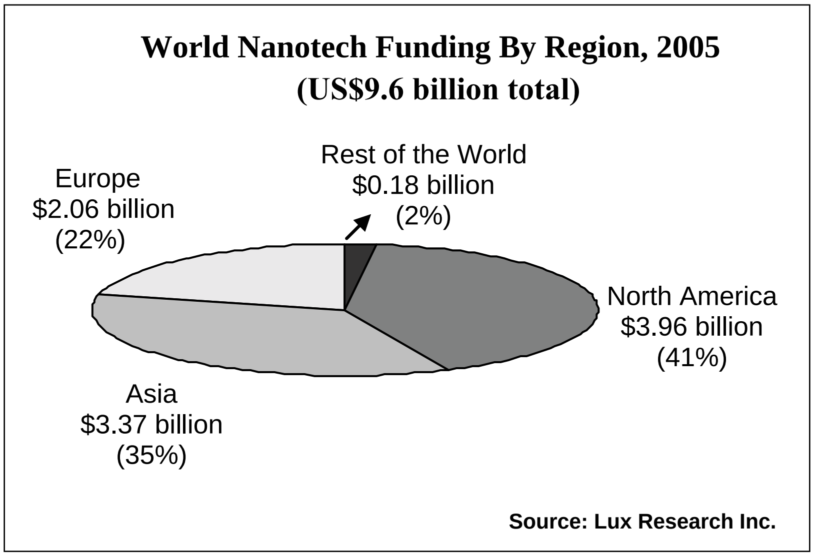


Figure 6. Results of research conducted by Lux Research Inc.

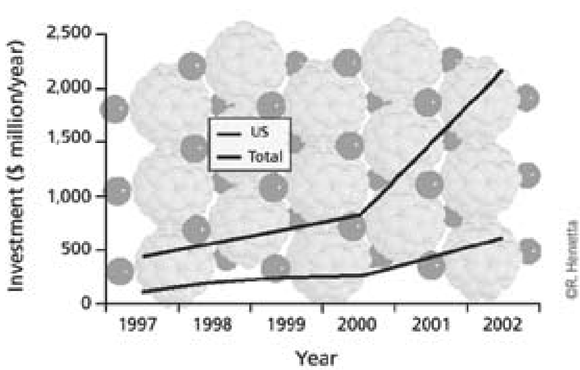


New analysis from *Frost & Sullivan*[13] (Figure 7), World Nanobiotechnology Market, reveals that sales of nanomaterials for use in *nanobiotechnology* applications generated revenues of *USD 750.0 million in 2004*. This figure is projected to reach *USD 2,056.5 million in 2011*.

U.S. National Science and Technology Council[14] estimated nanotechnology products to be around $1 trillion/year by 2015 with a breakdown in electronics and communication of $300 billion, materials and processing of $340 billion, life sciences of $180 billion, and sensors and instrumentation of $22 billion.

Analyzing the data from all the three researches mentioned above we can say that the future of bionanotechnology is extremely bright and flourishing.

Figure 7. Analysis from Frost and Sullivan[13]



CONCLUSION

Nanotechnology shows an extremely enormous potential for phenomenal works in the fields of medicine and biotechnology. In future, nanotechnology can help in the early and effective diagnosis of various diseases like cancer, Alzheimer, and so on.

The most significant applications of bionanotechnology lie in two major fields of therapeutics and diagnostics. We have discussed above some of the most prominent and researched applications of bionanotechnology like dendimers, personalized medicines, bionanosensors, regenerative medicines, quantum dots, gold nanoshells, silver nanoparticles, smart drugs, and so on.

Apart from studying the applications of bionanotechnology, we have also shown the future prospects of bionano. Currently, bionanotechnology is in the phase of being discovered and researched upon. In spite of its huge number of applications and being funded generously, it has very few products available in the market as of today. However, with the next five to 10 years, it is predicted that around 200 products based upon bionanotechnology will be FDA approved.

REFERENCES

Baba, Y. (2006a). Diagnosis of diseases by nanodevice. *Nippon Rinsho*, 271–277.

Baba, Y. (2006b). Nanotechnology in medicine. *Nippon Rinsho*, 189–198.

Chan, W.C. (2006). Bionanotechnology progress and advances. *Biol Blood Marrow Transplant*, 87–91.

Clark, J., Singer, E.M., Korns, D.R., & Smith, S.S. (2004). Design and analysis of nanoscale bioassemblies. *Biotechniques*, 992–6, 998–1001.

Connelly, M.A. *Nanotechnology—Its promises for the field of medicine and the* ethical implications of these promises.

Deepak. *The promise of nanobiotechnology.* Retrieved February 5, 2008, from http://techbizmedia.com/mn-doci/technology/the-promise-of-nanobiotechnology/

Emerich, D.F., & Thanos, C.G. (2003). Nanotechnology and medicine. *Expert Opin Biol Ther*, 655–663.

Emerich, D.F. (2005). Nanomedicine—Prospective therapeutic and diagnostic Applications. *Expert Opin Biol Ther*, 1–5.

ETC Group. (2006). *Nanotech Rx: Medical applications of nano-scale technologies: What impact on marginalized communities?* Retrieved February 5, 2008, from http://www.etcgroup.org/en/

Gerhard, F. (2002). *Bio-nanotechnology in Japan-venture capital, public policy & results*. Stanford University.

Grodzinski, P., Silver, M., & Molnar, L.K. (2006). Nanotechnology for cancer diagnostics: promises and challenges *Expert Rev Mol Diagn*, 307–318.

Ha, T.H., Jeong, J.Y., & Chung, B.H. (2005). *Immobilization of hexa-arginine tagged esterase onto carboxylated gold nanoparticles. Chem Commun (Camb)*, 3959–61.

Jain, K.K. (2005). Nanotechnology in clinical laboratory diagnostics. *Clin Chim Acta*, 37–54.

Jiang, W., Papa, E., Fischer, H., Mardyani, S., & Chan, W.C. (2004). Semiconductor quantum dots as contrast agents for whole animal imaging. *Trends Biotechnol.*, 607–609.

Kohli, P., & Martin, C.R. (2005). Smart nanotubes for biotechnology. *Curr Pharm Biotechnol*, 35–47.

Kubik, T., Bogunia-Kubik, K., & Sugisaka, M. (2005). Nanotechnology on duty in medical applications. *Curr Pharm Biotechnol*, 17–33.

Kvennefors, A., & Persson, F. (2004). *The technology and applications of bionanosensors.*

Lambkin, I. (2005). *Nanobiotechnology background paper.*

Lin, H., & Datar, R.H. (2006). Medical applications of nanotechnology. *Natl Med J India*, 27–32.

Moghimi, S.M., Hunter, A.C., & Murray, J.C. (2005). Nanomedicine: Current status and future prospects. *FASEB J*, 311–330.

Nishiyama, N., & Kataoka, K. *Nano-engineering for biomedical applications*.

Peppas, N.A. *Smart drug delivery and bionanotechnology*.

Rajangam, K., Behanna, H.A., Hui, M.J., Han, X., Hulvat, J.F., Lomasney, J.W., & Stupp, S.I. (2006). Heparin binding nanostructures to promote growth of blood vessels. *Nano Lett*, 2086–2090.

Svenson, S., & Tomalia, D.A. (2005). Dendrimers in biomedical applications--reflections on the field. *Adv Drug Deliv Rev.*, 2106–2129.

Weng, J., & Ren, J. (2006). Luminescent quantum dots: A very attractive and promising tool in biomedicine. *Cur Med Chem.*, 897–909.

Yamato, M., & Okano, T. (2006). Nanotechnology-based regenerative medicine—Cell sheet engineering utilizing temperature-responsive culture dishes. *Nippon Rinsho*, 335–341.

KEY TERMS

Bionanotechnology: Area of study where nanotechnology has applications in the field of biology and medical sciences.

Nanodiagonisis: Application of bionanotechnology for detection of diseases at molecular level.

Nanotechnology: Engineering of functional systems at the molecular scale.

Nanotherapeutics: Application of bionanotechnology in which the diseases can be treated by the use of various techniques at molecular level.

Nanoshell: A spherical core consisting of a particular compound, which is surrounded by a shell of a few nanometer of thickness.

Personalized Medicine or Patient-Specific Medicine: The prescription of specific therapeutics best suited for an individual.

Quantum Dots: Nanoparticles made out of semiconductors having unique electrical and optical properties.

Regenerative Medicine: Helps in natural healing processes to work faster, or uses special materials to regrow missing or damaged tissues.

D

Dealing with the Primacy of Knowledge in an In-Patient Mental Health Setting

Ivor Perry
de Montfort University, UK

INTRODUCTION

This article brings together three strands of interest to IS researchers: the problem of managing knowledge, particularly deep experiential knowledge; the opportunities and limitations of process and process-based systems (PBS); and the interplay of organizational culture across both. Similar work has been done in medically oriented departments, such as pediatrics and admissions, where a link between PBS and changes in organizational culture has been observed (Perry, 2003) and where an argument has been developed that PBS actively contribute to the creation of organizational knowledge (Perry, 2004).

Healthcare processes can be complex, and often appear to be unique to the patient (Smith, 2000). Although case mix systems and, in the last few years, electronic patient record (EPR) systems, have become common (Smith, 2000), the problem of handling highly iterative, possibly idiosyncratic processes remains. Although "care pathways" can be defined and often executed with little change, many treatments nevertheless are composed largely of ad hoc patient-carer interactions where the carer may be using a large amount of experience and expertise, overlaying their training (Ward, 1992).

A mental health unit was selected as the setting for this particular piece of research because it displays characteristics that are qualitatively different from those other areas of health care (Altschul, 1997). Mental health units, due to the nature of their work, deal with a myriad of variables in terms of physical, behavioral, and cognitive disorders, and staff appear at least to use a high degree of tacit knowledge (Clarke, 1999).

The results, it will be argued, point toward the value of PBS rather than the more static use of databases. On the other hand, the problem of interpreting system-generated data, unaccompanied by interpretive clues from other members of the community of practice, still remains.

Kakabadse, Kakabadse, and Kouzmin (2003) discuss a continuum of types of knowledge that runs through data-information-realization-action-wisdom. Substantially, they identify a deepening complexity and usefulness of knowledge as more and more cognitive and interpretive processes are deployed. At the far end of the continuum, wisdom therefore becomes a mode of symbolic processing by a highly developed will (Kakabadse et al., 2003) Importantly, intention and personality, including life experience, are essential to make the most use of knowledge. It is this aspect of knowledge management that forms a key issue to be dealt with in this article.

PROCESS-BASED SYSTEMS

PBS can be defined broadly as systems that enact business processes. They may be highly prescriptive, like some of the insurance claims applications in large financial institutions, where agents have little choice in the order and number of steps they must follow. They may be ad hoc systems, like some groupware applications. They may be what has been described as proto-PBS (Perry, 2004), systems such as electronic patient records that do not automatically move from one stage to another but indicate when a new stage may be necessary (e.g., viewing the results of a hematology report).

Additionally, the EPR records themselves constitute the content of a number of organizational processes, particularly in the area of patient care. Systems that *address processes rather than transactions* have been an area of research focus for some time (Doherty & Perry, 1999). The argument has been that these systems accrete organizational knowledge as they move through their process; in this way, they help to create and expand organizational knowledge by bringing existing and new data together and prompting staff to look afresh at the available information (Perry, 2004).

An example of a proto-PBS in a mental health setting is discussed by Bloomfield (2005), who describes the electronic whiteboard used by a number of District Health Boards (DHBs) in New Zealand. This system—officially termed the Current Customer View (CCV)—acts as a metalayer for existing mental health systems. It permits staff at a DHB to view important patient data that may be held on the systems of other DHBs. This is significant, as mental health patients may well move abruptly and without informing their doctor; in such situations, clinicians in a new area need current information about the patient's history, recent problems, behavior, and medication (Bloomfield, 2005).

Managing knowledge in health care

A general problem in the discussion of knowledge management in any complex social setting is to understand the nature of the knowledge being created, manipulated, or communicated. Polanyi (1958) was among the first to identify the significance of tacit knowledge, while Nonaka (1990) reinterpreted the concept, asserting that tacit knowledge could be made explicit. Hitt and Tyler (1991), studying the behavior of business executives, concluded that they use automatic and nonconscious processes. Further, they believed that these processes drew upon experientially established cognitive structures (Hitt & Tyler, 1991). We can see how significant tacit knowledge can be; these processes and structures were being used in making *strategic business decisions*. Later, Nonaka (1994) noted how people used stories, metaphors, and analogies in order to share and communicate tacit knowledge. This sharing of oblique or interpreted knowledge is central to the problem of complex knowledge management, since health care includes not only medical evidence but also opinion and experience, particularly in relation to patient care (Priebe & Slade, 2002).

CULTURE AND CLIMATE

Organizational culture is most often revealed in how it and the organization are perceived by organization members (Schein, 1999). In large health care organizations and even within hospitals, it is to be expected that a variety of cultures will exist (Helms & Stren, 2001; Hofstede, Neuijen, Ohayv & Sanders, 1990). Additionally, research seems to show that whatever the type of culture, it will affect the perceptions of the staff. Similarly, major organizational changes will affect staff attitudes and practices, and this will affect culture (Schein, 1999). In the NHS (the United Kingdom's National Health Service), a study during intense change was carried out by Litwinenko and Cooper (1994). They found not only that the proposed changes cut across traditional beliefs and expectations held by the health care worker, but also that all types of culture present in the organization were changing (Litwinenko & Cooper, 1994).

A critical element of nursing practice and knowledge transfer is 'organizational climate'. (R. B. Brown & Brooks, 2002) This is associated with but different from organizational culture, which is the broad set of commonly held attitudes, beliefs, and assumptions that characterize an organization (Brown, 1998; Sathe, 1985; Schein, 1999; Smircich, 1983). Organizational climate refers to the atmosphere that employees perceive in their organizations. It is created by practices, procedures, and rewards, and may differ markedly from one hospital department to another. Organizational climate is also associated with shared emotion or feelings. This emotional dimension has been shown to be both a social influence on the behavior of individual staff members and on their collective actions (Brown & Brooks, 2002). As a result, Brooks and Brown identify emotional climate as an important social construct in the interaction among staff. While their study does not specifically address knowledge management or creation, it does make it very clear that individual and group self-identification both stem largely from stories and from shared experience. For students of knowledge transfer, this observation is highly significant. In a nursing environment, we have to consider not just the transfer of explicit knowledge or the passing on of know-how, as defined by Nonaka (1991). We are also dealing with feelings and emotions, which are sometimes very deep-seated and contribute to the identity of the unit.

MANAGING KNOWLEDGE IN MENTAL HEALTH

Scott (2005), in an informative paper on knowledge workers as part of a social network, makes some initial points: "Their daily work may be unpredictable, multidisciplinary, and nonrepetitive. The jobs assigned to them have long term goals and, due to the relatively standard ambiguity and complexity of the task, require knowledge workers to routinely collaborate with co-workers to utilize multiple viewpoints to solve problems" (Scott, 2005). This accords well with the observed pattern of work in mental health care; standard mental health nursing texts refer frequently to the need to act as a team, to discuss findings with colleagues, to prepare handovers for other team members as they come on shift, and to discuss issues with patients both within and outside of formal interviews (Ward, 1992).

This raises the question not only of interstaff knowledge transfer, but also of staff-patient knowledge transfer in a mental health setting. Traditional, authority-based approaches to psychiatry changed markedly from the 1960s (Clarke, 1999). The nonjudgmental counseling of Carl Rogers emphasized the acceptance of the client's behavior and attitudes, underscoring the essential place that tacit knowledge, in its broad sense, occupies in mental health care. The work of rigorously postmodern psychiatrists such as R.D. Laing in the 1960s threw the central position of tacit knowledge into sharp relief. Effectively, he deconstructed traditional medical discourse and offered an alternative vision where indeterminacy and paradox created unexpected but liberating states of mind (Clarke, 1999).

Laing's approach to psychiatric knowledge was not dissimilar to Foucault's conviction that knowledge, or rather the socialized use of knowledge, is a primary source of power. Like Foucault, Laing was of the opinion that the construction of knowledge can be observed in the discourse that is employed by social groups, and that the discourse of the powerful can then be deconstructed to provide alternative interpretations of reality that will effectively give power to the disenfranchised (Butler, 2002; Seidman, 1998).

Meanwhile, mainstream mental health care has become increasingly evidence-based since the 1980s, although this should not be taken to mean that all evidence is simplistically statistical in nature (James & Burns, 2002). On the one hand, evidence-based medicine has been and is being used to demonstrate cost efficiencies and savings (James & Burns, 2002). On the other hand, the postmodern legacy has been to challenge the objectivity of scientific medicine and to highlight the evidence of patients. Laugharne (2002) contends that although clinicians are educated in the scientific method, the practice of mental health care should not be a scientific exercise but an exercise in humanity, informed by ethical and moral choices.

If evidence-based practice is being seriously questioned (or at least the type of evidence is being seriously questioned) in mental health practice, where does that leave the issue of knowledge management in this area? Unsurprisingly, perhaps, researchers such as Allard (2002) believe that progress in this area is inextricably linked with discovering the hidden assumptions and agendas at play in mental health research. His call is for evidence that takes into account the value systems of users/survivors (of mental health care) (Allard, 2002).

METHODOLOGY

Knowledge management is studied seriously within the health care environment, but it appears that comparatively little is discussed in mental health care, which appeared to use a large element of tacit knowledge. Also (speculatively) perhaps *because* other branches of health care can benefit so quickly and directly from the introduction of explicit knowledge management, general medicine and explicit knowledge handling have occupied a good deal of attention in the literature.

In order to form some understanding of the relationship (if any) between the available literature and the actual practice and perceptions of practitioners in mental health, it was decided to follow Churchill's (1991) approach and undertake an exploratory study. A specific choice was therefore made to begin with a discrete unit. A mental health unit within a large general hospital in the United Kingdom was selected, and nursing staff was interviewed at intervals over a three-month period.

In order to obtain a broad perspective and to elicit staff's own views, it was decided to employ a mixed methods approach in order to provide a degree of triangulation, as described by Flick (2004) and Hildebrand (2004). Central to this concept was the framework of grounded theory (Hildebrand, 2004) on several grounds. First, it offers a conceptual framework in which multiple data streams from literature, observation, and narrative can be drawn together through a recursive sense-making process. Second, it engages the contribution of practitioners and their perspectives, for example, through both ethnographic and phenomenological studies. Finally, it allows space for the interplay between researcher and practitioner in order to develop new, independent perspectives (Leonard & McAdam, 2001).

Within a longitudinal timeframe, as recommended by grounded theory concepts, the interviews were considered and assessed using a modified form of discourse theory. The exploratory nature of the research, for example, did not permit extended time to revisit conversations more than once in order to gain multiple and deeper insights into the emotional cognition of the actors in the mental health unit (Mangham, 1998). On the other hand, some techniques of discourse theory proved very useful (e.g., to allow an interviewee to feel relaxed so that when he or she was describing his or her world, his or her perception was as useful as the worlds being described by others. Further, discourse theory allows the researcher to examine the positioning of himself or herself as interpreter, and of the interviewee as storyteller. Both these positions are relevant to knowledge transfer and to knowledge creation, as they form the building blocks of mixed explicit and tacit knowledge (Davies & Harre, 2001).

Throughout the interviews, cognizance was taken where possible of nonverbal communication in order to compare with the language dimension offered by discourse analysis. Within this area, attempts were made to compare nonverbal styles with task, authority, and gender roles (Remland, 2000).

RESULTS

One of the most remarkable findings was the strength of feeling (i.e., emotional climate) surrounding the awareness of group identity. Comments such as "Mental health nursing isn't like anything else—even in the NHS, they don't realize how different it is" were usually accompanied with nonverbal stresses of hand or eye movements, or raised voices.

Nurses also identify with the patients as much as with the concept of being clinical staff. "When someone says they're depressed, we know what they mean. There's no point in telling them that they're not depressed, because that's a particular medical diagnosis, and they're suffering from something else." It appears that, mirroring de Pinheiro and Spink's (2004) findings, meanings are being sought and negotiated to avoid the therapeutic relationship becoming trapped into technical rationality.

It became evident that this unit, so far as the nursing staff were concerned, were functioning as a community of practice (Wenger, 1998): "Coffee breaks and handovers are very important. We talk about events and patients, but we're also saying to each other, 'I know what it's like; we're all in the same boat'"; "The way we work is based on best practice—yes, obviously—but we have our own ways of just signaling to each other that a certain patient is having a difficult day, or that a particular doctor is making some daft decisions." The interview responses offered all three of Wenger's community of practice dimensions: (1) nurses saw themselves as engaging in a joint enterprise understood and *continually renegotiated* by the members [emphasis added]; (2) the mutual engagement had bound members (interestingly, patients seemed included, too) into a social entity; and (3) it had developed a *shared repertoire* of communal resources (routines, sensibilities, artifacts, etc.) (Wenger, 1998).

The relevance to knowledge management of this sort of community of practice is quite profound and touches at the heart of the problem that sooner or later information systems must deal with. The group—in fact, the human process that the group enacts—retains knowledge in living ways, unlike a database or manual (Wenger, 1998).

The ways in which knowledge is currently recorded are varied and regarded in different ways. At the moment, the unit has access to the hospital's patient administration system (PAS), although this is regarded at best with a degree of condescension. It is criticized as being often inaccurate and out of date. Therefore, it is rarely relied upon, and staff uses corroborating data when they can, usually from the patients themselves. Much of the useful (i.e., according to the nurses) data are recorded on paper within the unit itself. On the one hand, the PAS is held in low esteem, but on the other hand, computing technology is seen to be advancing rapidly. Again, (perhaps defensively) staff questions whether any form of technology can adequately capture the kind of knowledge they need to impart to colleagues.

DISCUSSION

Although there is comparatively little written on the subject of knowledge management in mental health, there is no shortage of evidence. It is the nature of that evidence that poses problems for the researcher.

Strictly speaking, information systems need not be technology- or computer-driven at all. Most management information systems textbooks will offer a definition that allows for the nonuse or the partial use of technology (O'Brien & Marakas, 2006).

The richness of the evidence surrounding the types of knowledge and the physical expressions of knowledge (e.g., individual behavior, cognitive processes, work processes, PAS, handwritten notes, etc.) should not be seen as a problem for the IS researcher, but rather the beginning of the solution. Blackburn's principle of complementarity (described by Gray)—full understanding of a phenomenon requires the application of a number of complementary theories or descriptions (Gray, 2000)—helps us move forward by recognizing that an array of interpretations of knowledge is probably essential in finding better ways of creating, managing, and transferring it. Information systems must be flexible enough to support these interpretations.

The current consumer view (CCV) used in New Zealand (see Introduction) is very much like a proto-PBS (process-based system) (Perry, 2004) that shares similarities with groupware systems (Ciborra, 1996) in terms of initiating processes, even if those processes are then continued outside the information system itself. The CCV is also significant in knowledge management terms, as it provides the interface between the data management world and the human process world of cognate functions, know-how, and emotions.

Since most other means of handling knowledge in mental health appear to be part of a process or initiate processes, it would appear that the use of technology-based systems with a process capability would be much closer to the way that staff members seem to work. In other words, just as the CCV initiates and, to a limited extent, enacts a process, so IS professionals might start to embed process within their systems wherever possible, not in a prescriptive way, but in a permissive sense that enables the staff to select and use process rather than just extract or load data.

CONCLUSION

This article has explored the interplay between explicit knowledge and the extensive, often hidden environment of tacit knowledge. It has identified the importance of these issues in mental health care as a work environment that utilizes and depends on such tacit knowledge to a greater degree than medical nursing or many other professions. In doing so, it continues a perspective on knowledge management that has been accepted for a number of years already—we should consider the topic of information systems as more of a social science or a sociological subject, not simply a technical one (Galliers, 1993).

While many vendors (CSE-Servelec, 2006; Graphnet Health Ltd, 2006; iSoft plc, 2006) are working on or have produced case work software for mental health practice, all have stopped short of consciously harnessing the potential of the deeply held, frequently unconscious knowledge of nursing staff. The commercial judgments of the vendors are, of course, outside the scope of this article, but it is worth looking at the problems they would face since they are largely the same problems that the IS researcher faces when considering the role of technology-based systems in mental health care.

First, there is the problem of identifying what the tacit knowledge is. Looking at the evidence so far, it is most frequently observed in behavior, both linguistic and nonverbal. It is grounded in the organizational culture, which, as we have seen, may be quite specific in a small, tightly identified or geographically isolated unit (Helms & Stren, 2001). It is also grounded in the emotional climate of that culture, which may be even more specific (Brown & Brooks, 2002).

Second, there are the technological problems of capture, storage, and communication. Some of the case work systems (e.g., Graphnet) routinely use large data objects (X-ray images, scanned document images, etc.) and employ technological standards like XML to capture, sort, and display a variety of data from many different formats.

Speculatively, it may be viable for staff members to record their data as transcript and use video/voice recognition software to encode the data as text; then it can be used easily within conventional systems like EPR. This would leave the video/voice recording as an indicator for colleagues of the emotional content. On the other hand, such systems would perhaps be so far outside the established norms of information requirements that it would be at direct variance with organizational culture, which, as researchers have noted for many years, would result in very strong user resistance, quite likely leading to system failure (Cooper, 1994).

There is also the argument, proposed by a number of psychiatrists and mental health researchers, that the very act of fitting data into neat classifications is problematic when trying to treat patients with mental illnesses. Without going to the extremes of Rogers or Laing, such ideas are now becoming part of mainstream thinking. The more one takes account of human experience, the more difficult it is to utilize the language of diagnostics and therapeutics (Clarke, 1999). The argument even can be made that database-type classifications have dubious authority as knowledge, since each patient has to be viewed as a unique entity.

If knowledge in mental health care is heavily dependent on linguistics and behavior, then what role can information systems realistically play? Should they simply be relegated (as they often are, and as the PAS in this study was) to the storage and retrieval of objective data such as name, date of birth, and address? Should mental health information systems instead be transformed from their (positivist, medical) linguistic basis into (subjectivist) semiotics-based systems? Probably not.

First, information systems developers, in conjunction with users, need to be aware that far more knowledge is at work than currently can be managed technologically. Second, both parties should be aware as they design systems that, especially in mental health, the medically-sanctioned terminology and process will exercise power both of action and of meaning over staff and patients alike.

Second, there is much that can be done to improve existing systems. Whereas the PAS in this study was effectively a straightforward transactional system (Perry, 2003) and therefore did little to support nursing processes, process-based systems like the CCV (Bloomfield, 2005; Orr, 2005) can win considerable enthusiasm by demonstrating that they can act as an expert assistant to the practitioner, to the point of becoming a virtual colleague. As Maier and Remus (2003) point out, such process knowledge management (PKM) systems will, at a functional level, include personalization, contextualization, *navigation from knowledge elements to people* [emphasis added]. There is evidence that process-based systems enable organizational functions to run more easily, conveying not just data, but knowledge, thus changing the normative expectations of users (Perry, 2004). This is in line with Tsoukas' (1996) suggestion that the use of knowledge is conditioned by the normative expectation of people's organizational function.

To return to the paradigm of data-information-realization-action-wisdom (Kakabadse et al., 2003), if wisdom is the highest expression of knowledge and is exercised by the operation of a purposeful will, then successful mental health interventions (by the nature of the knowledge exercised) require wisdom and all the supporting dimensions of knowledge. The evidence seems to point to process-based systems, whether they are EPR, case work, or New Zealand's CCV, to be the most effective means of joining the tacit interpersonal dimension of practice with the explicit, medical, and procedural dimension.

REFERENCES

Allard, S. (2002). A user/survivor perspective: What's behind the evidence? In M. Priebe (Ed.), *Evidence in mental health care* (pp. 207-215). Hove: Brunner-Routledge.

Altschul, A. (1997). A personal view of psychiatric nursing. In S. Tilley (Ed.), *The mental health Nurse: views of practice and education*. Oxford: Blackwell Science.

Bloomfield, S. (2005). Advancing mental health pathways. *Health Care and Informatics Review Online*.

Brown, A. (1998). *Organisational culture* (2nd ed.). Harlow, England: Prentice Hall.

Brown, R.B., & Brooks, I. (2002). Emotion at work. *Journal of Management in Medicine, 16*(5), 327–344.

Butler, C. (2002). *Postmodernism: A very short introduction*. Oxford: Oxford University Press.

Churchill, G.A. (1991). *Marketing research methodological foundations* (5th Ed.). Orlando, FL: Dryden Press.

Ciborra, E. (1996). Groupware and teamwork: Invisible aid or unnecessary hindrance? Chichester, England: Wiley.

Clarke, L. (1999). *Challenging ideas in psychiatric nursing*. London: Routledge.

Cooper, R.B. (1994). The inertial impact of culture on IT implementation. *Information and Management, 27*, 17–31.

CSE-Servelec. Retrieved February 28, 2006, from http://www.cse-servelec.com/

Davies, B., & Harre, R. (2001). Positioning: The discursive production of selves. In M. Wetherell, S. Taylor,

& S.J. Yates (Eds.), *Discourse theory and practice*. London: Sage.

de Pinheiro, O.G., & Spink, M.J.P. (2004). Discursive practices and demographic participation: Negotiating language in mental health. *Journal of Health Psychology, 9*(1), 55–71.

Doherty, N.F., & Perry, I.F. (1999). The uptake and application of workflow management systems in the UK financial services sector. *Journal of Information Technology, 14*(2), 149–160.

Flick, U. (2004). Triangulation in qualitative research. In U. Flick, E. Von Kardoff, & I. Steinke (Eds.), *A companion to qualitative research* (pp. 178–183). London: Sage.

Galliers, R.D. (1993). Research issues in information systems. *Journal of Information Technology, 8*, 92–98.

Graphnet Health Ltd. Retrieved February 28, 2006, from www.graphnethealth.com

Gray, P.H. (2000). The effects of knowledge management systems on emergent teams: Towards a research model. *Journal of Strategic Information Systems, 9*(2–3), 175–191.

Helms, M., & Stren, R. (2001). Exploring the factors that influence employees' perceptions of their organisation's culture. *Journal of Management in Medicine, 15*(6), 415–429.

Hildebrand, B. (2004). Anselm Strauss. In U. Flick, E. Von Kardoff, & I. Steinke (Eds.), *A companion to qualitative research* (pp. 17–23). London: Sage.

Hitt, M., & Tyler, B. (1991). Strategic decision models: Integrating different perspectives. *Strategic Management Journal, 12*, 327–352.

Hofstede, G., Neuijen, B., Ohayv, D.D., & Sanders, G. (1990). Measuring organisational cultures: Implications for training and development. *Administrative Science Quarterly, 35*, 286–316.

iSoft plc. Retrieved February 28, 2006, from www.isoftplc.com

James, P., & Burns, T. (2002). Influence of evidence on mental health care developments since 1980. In M. Priebe (Ed.), *Evidence in mental health care* (1st ed.) (pp. 28–40). Hove: Brunner-Routledge.

Kakabadse, N.K., Kakabadse, A., & Kouzmin, A. (2003). Reviewing the knowledge management literature: Towards a taxonomy. *Journal of Knowledge Management, 7*(4), 75–91.

Laugharne, R. (2002). Evidence—The postmodern perspective. In M. Priebe (Ed.), *Evidence in mental health care* (pp. 53–57). Hove: Brunner-Routledge.

Leonard, D., & McAdam, R. (2001). Grounded theory methodology and practitioner reflexivity in TQM research. *International Journal of Quality and Reliability Management, 18*(2), 180–194.

Litwinenko, A., & Cooper, C.L. (1994). The impact of trust status on corporate culture. *Journal of Management in Medicine, 8*(4), 8–17.

Maier, R., & Remus, U. (2003). Implementing process-oriented knowledge management strategies. *Journal of Knowledge Management, 7*(4), 62–74.

Mangham, I.L. (1998). Emotional discourse in organizations. In D. Grant, G. Keenoy, & C. Oswick (Eds.), *Discourse + organisation* (pp. 51–64). London: Sage.

Nonaka, I. (1990). Redundant, overlapping organization: A Japanese approach to managing the innovation process. *California Management Review, 32*(3).

Nonaka, I. (1991). The knowledge creating company. *Harvard Business Review* (Nov-Dec), 96–104.

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. *Organization Science, 5*, 4–37.

O'Brien, J.A., & Marakas, G.M. (2006). *Management information systems*. New York: McGraw-Hill.

Orr, M. (2005). Knowledge management to systemic wisdom development. *Health Care and Informatics Online* (September 2005).

Perry, I.F. (2003). Making sense of the organisation's knowledge: Does systematisation of the knowledge base have a positive or negative effect on organizational culture? In *Proceedings of the HICSS '36* (36th Hawaii International Conference on System Sciences), Hawaii.

Perry, I.F. (2004). Knowledge as process, not data: The role of process based systems in developing organisational knowledge and behaviour. In *Proceedings of the Americas Conference on Information Systems (AMCIS-10),* New York.

Polanyi, M. (1958). *Personal knowledge: Towards a post-critical philosophy*. London: Routledge and Kegan Paul.

Priebe, S., & Slade, M. (2002). Evidence in the twenty-first century: The way forward. In M. Priebe (Ed.), *Evidence in mental health care*. Hove: Brunner-Routledge.

Remland, M.S. (2000). *Nonverbal communication in everyday life* (1st ed.). Boston: Houghton Mifflin.

Sathe, V. (1985). *Culture and related corporate realities*. Homewood, IL: Richmond and Irwin.

Schein, E. (1999). The corporate culture survival guide: Sense and nonsense about culture ch. San Francisco: Jossey-Bass.

Scott, P.B. (2005). Knowledge workers: Social, task, and semantic network analysis. *Corporate Communications: An International Journal, 10*(3), 257–277.

Seidman, S. (1998). *Contested knowledge: Social theory in the postmodern era* (2nd ed.). Oxford: Blackwell Publishers Ltd.

Smircich, L. (1983). Concepts of culture and organisational analysis. *Administrative Science Quarterly, 28*.

Smith, J. (2000). *Health management information systems*. Buckingham: Open University Press.

Tsoukas, H. (1996). The firm as a distributed knowledge system: A constructionist approach. *Strategic Management Journal, 17*(1), 11–25.

Ward, M.F. (1992). *The nursing process in psychiatry* (2nd ed.). Edinburgh: Churchill Livingstone.

Wenger, E. (1998). Communities of practice: Learning as a social system. *Systems Thinker,* (June).

KEY TERMS

Evidence-Based Practice: The use of external evidence, including random trials, rather than a reliance on clinical authority.

Process-Based Systems: Systems that encapsulate or enact business processes. Examples may be workflow systems, CRM systems, groupware, electronic patient records.

Organizational Climate: The atmosphere that employees perceive in their organizations. It is created by practices, procedures, and rewards.

Organizational Culture: The broad set of commonly held attitudes, beliefs, and assumptions that characterize an organization.

A Decision Based Support System Based on GIS Technology

Vito Albino
Politecnico di Bari, Italy

Giuseppe Digregorio
Politecnico di Bari, Italy

Guido Sechi
Politecnico di Bari, Italy

INTRODUCTION

The continuous increase of the national expense for the healthcare services, the necessity to adjust the performance of the service of the National Healthcare System to the change of the demand of health coming from the population, and the necessity to evolve towards a logic of integration along a *"continuum of care,"* push for the reorganization of healthcare services inside the local healthcare district. This should permit:

1. The increase of the efficiency level, by means of economies of scale, know-how diffusion, the elimination of redundant resources;
2. The acquisition of competitive advantage for the network components, which are focused on specific areas and are subject to a smaller competitive pressure, due to the existence of protection mechanisms;
3. The increase of clinical effectiveness; and
4. The increase of the quality of service.

On this basis, it is possible to define the main factors to take into account for the reorganization of healthcare services inside the healthcare district: (1) the territorial distribution of the services; (2) the citizen-customer satisfaction, in terms of expectations and preferences; (3) the centralization of high complexity services as a guarantee of quality and cost sustainability; (4) the appropriateness of the healthcare services, in relationship with the local needs of health.

The study of these factors implies the analysis of a great amount of quantitative and qualitative variables, and the management of different types of data, including geographical information.

According to Hopkins (1977) and Collins et al. (2001), such a problem can be expressed as a complex localization problem, and consists of the identification of the most suitable location, according to specific priorities and to the features of the activities to be performed.

In particular, its application to the integrated healthcare network model consists of the research of the optimal location for local healthcare services.

A useful tool to support the solution of localization problems is represented by Geographical Information Systems (GISs) (Mullner, Chung, Croke, & Mensah, 2004). Such technologies are able to store, process, and represent large quantities of data, combining data management and spatial referencing capabilities. They are also able to screen the set of alternatives, and so, they can provide support when the set of location alternatives is very large (Eastman, Kyem, & Toledano, 1993). Because of this, they can be used as support tools for solving optimisation problems. This can be done by integrating optimisation models in a GIS environment (Lin & Kao, 1998). However, the growing dimension of the problem makes the integration of optimisation techniques into a GIS environment very difficult, increasing the computational complexity.

The application of *multiattribute* and *multicriteria methods* (Chan, 2001), based on heuristic techniques, can be an alternative solution. These methods cannot lead to the optimal solution, but they permit to identify "most suitable" locations, even when the complexity of the problem is high (Malczewski, 2004). They are generally based on criteria weighting techniques, such as Analytic Hierarchy Process (Saaty, 1990) and Simple Additive Weighting (Carver, 1991; Eastman et al., 1993), and their use is suggested when technological limitations do not permit to use more complex methods to calculate the optimal solution (Cromley & Hanink, 1999).

LOCALIZATION METHODOLOGY

The solution of the localization problem of a healthcare network cannot be disjointed from the analysis of all those factors influencing the perception of the health services by the patients. Such factors have both a quantitative dimension (e.g., the distance from the essential service), and a qualitative one (e.g., the perceived quality level of local healthcare services, or the chance to choose the place where to receive assistance).

With reference to this, Fortney, Rost, and Warren (2003) point out five possible dimensions for the analysis: affordability, acceptability, accomodation, availability, and accessibility. The last one tends to be extremely useful for examining the spatial interaction between healthcare services and patients. Nevertheless, the accessibility problem has to be solved, referring to the analysis of the demographic, economic, and social characteristics of the healthcare district, aimed at identifying the health problems and needs of the population (Curtis & Taket, 1996) and pointing out possible situations in which the risk to be not cured, or to be cured late, is very high (Brown, 1988).

According to Harper, Shahani, Gallagher, and Bowie (2004) the localization of the healthcare services is based on eight factors considered relevant for spatial analysis:

1. **Services and clinical speciality:** Analysis of the currently provided services, in terms both of medical, surgical, or diagnostic services, and of resources (namely, number of beds, physicians, nurses, and so on).
2. **Actual location of the healthcare centres:** Analysis of the localization in the examined area of the services with high and low specialization, aimed to identify zones in which it is necessary to locate new services or zones in which there is useless redundancy.
3. **Distribution of the population:** Analysis of the potential patients who reside in the area of the existing healthcare centres, also using social and economic data.
4. **Demand of health:** Analysis of the health demand referred to different healthcare services.
5. **Waiting times for the access to the cares:** It is important to find the trade-off between resources and system performance, in terms of waiting time for the access to cares.
6. **Routes:** It is necessary to assure the accessibility of all patients to given centres.
7. **Estimation of the minimal area:** It is opportune to verify if the potential demand of healthcare services, which characterizes a fixed area, is high enough to justify the location of a new healthcare centre.
8. **Patient preferences:** This concept can be expressed by means of a set of qualitative variables: for instance, the reputation of the healthcare centre, the presence of renowned experts (the patient is usually disposed to wait a long time for a high quality of service), and the perceived quality of services.

Aiming at finding a solution to this complex localization problem—with the specific objective to identify the most suitable location for a new healthcare service in the local healthcare district—a GIS-based methodology has been developed and proposed, adopting the following steps (Digregorio, 2006):

1. Collection of data related to the healthcare district area;
2. Definition of the criteria which affect the choice of the location for a new healthcare service. It is possible to identify two main categories:
    a. the admissibility criteria—the criteria which are considered essentials for eligibility;
    b. the optimality criteria—the criteria which are used for identifying the most suitable location between those who satisfy all admissibility criteria.

    At this step, the analysis considers only basic healthcare services (medicine), such as cardiology, paediatrics, audiology, geriatrics, neurology;
3. Identification of the admissible solutions—the solutions that satisfy the admissibility criteria;
4. Assignation of weights to optimality criteria and the application of the spatial analysis models for the search of the most suitable location (optimal solution)—that is, the admissible location

which receives the highest mark according to the weighted sum of the values for such criteria (Benabdallah & Wright, 1992; Lin & Kao, 1998);

5. Verification of variability and coherence of the optimal solution by means of sensitivity analysis, according to the variation of criteria relative weights;
6. Expanding of the procedure to the localization of surgical and diagnostic services. These services could present some different features from medicine that could affect the eligibility criteria and, consequently, the choice of the most suitable location.

CASE STUDY

A case study is proposed to show how the methodology can be used to investigate a healthcare district which includes 13 municipalities of the province of Bari, in southern Italy, with about 577,000 inhabitants. The existent healthcare services are analysed in detail by means of data provided by the two hospitals which are directly managed by the local sanitary authority. A database is created for demographic, social, economic, and geographical data, related to the spatial distributions of the existing hospitals and outstanding healthcare centres.

The next step consists of defining whether it is opportune to activate new healthcare services (in this case, cardiological services). The parameters which express the waiting time for the access to cardiologic cures are defined in order to measure the actual effectiveness of the service, and, on the basis of historical data, the aggregate rate of cardiovascular illnesses is assumed equal to 5% of the population.

Preliminarily, through the use of GIS tools, the healthcare district has been divided into parcels of area equal to one squared kilometer, characterized by the corresponding amount of potential patients.

Looking for the *admissible solutions* of the problem, the following criteria are chosen, and for each one, a corresponding map of suitable areas is created using a GIS:

a. **Potential patients/kmq:** The suitable areas are those where the density of potential patients is not too low. In particular, the areas which have a density equal or higher than the district average have been considered admissible.
b. **Competence areas:** Based on the current location of cardiology services, on queueing and performance indicators, the areas served by such units are identified: current unit performance and corresponding amount of potential patients are jointly considered. The areas showing an acceptable level of service are not included in the set of admissible solutions. This criterion aims at identifying the areas which particularly need the improvement of service.
c. **Proximity to routes:** The admissible areas are those which are located within a distance of one kilometer from main routes.

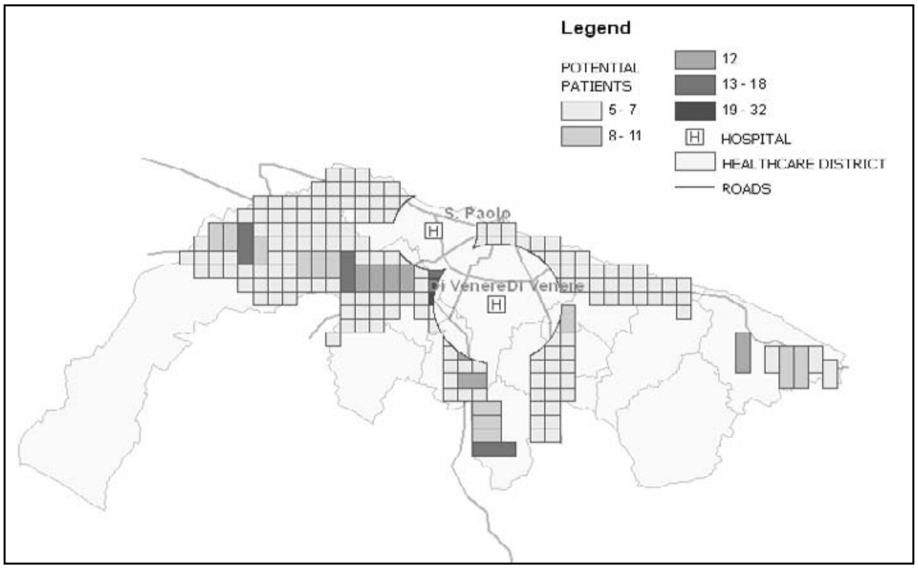
The set of admissible solutions is defined as the intersection of areas verifying the previously mentioned criteria (see Figure 1).

The next step consists of investigating the set of admissible solutions, looking for the *optimal solution* for the location of the new cardiological service. The afterwards procedure is based on the Simple Additive Weighting model (Carver, 1991).

Initially, the following criteria have been chosen:

d. **Distance from indifference areas:** This criterion considers that the part of the population which, according to its own geographical position, cannot choose between a set of healthcare service alternatives is disadvantaged in comparison with those who live within a small distance from more than one healthcare service. We define indifference areas as those within which citizens do not show a particular preference towards a healthcare centre, rather than others. According to such a criterion, the locations which are far from such areas are considered more suitable for locating a new healthcare service.
e. **Proximity to existing centres:** The criterion is based on a cost function which describes the opportunity to locate a new healthcare service where a healthcare centre without such service is already located.
f. **Per capita income:** The criterion is based on considerations of social equity, according to which it is fair to locate a new service in an area where the per capita income is low.

Figure 1. Admissible solutions



g. **Passive mobility of population:** Areas with high passive mobility (i.e., the tendency of inhabitants to use healthcare services located in other areas) are considered more suitable for locating a new healthcare service.

The weights to associate to the four criteria have been determined by applying the AHP methodology (Saaty, 1990), with the aim to express in quantitative information considerations concerning the relative importance of criteria. Relative importance between criteria x and y is assumed to be: 1 – when x and y are equally important; 3 – when x is more important than y; 5 – when x is much more important than y. The resulting optimal weight vector for the above criteria D, E, F, and G is the following:

$\boldsymbol{W^{opt}} = [W^{opt}_D ; W^{opt}_E; W^{opt}_F; W^{opt}_G] = [0.03; 0.32; 0.21; 0.44]$

Then, according to each criterion, the values for every location are calculated. Since they are expressed in different measure units, in order to make them comparable for being summed, they have been translated into a homogeneous scale of integer values between 0 and 8.

It has then been possible to identify the optimal solution, by means of the weighted sum of the values for the four criteria, calculated for each location (Figure 2) by means of GIS spatial analysis functions. The values for each criterion are represented by the map of values associated to possible locations. Obviously, only admissible solutions are taken into consideration. Among such locations, the one with the highest score was chosen as the most suitable area for the location of the new healthcare service.

As a final result, the opportunity of offering new services in an area where such services are not offered—with a low per capita income, well-served by the main routes, and affected by phenomena of population mobility towards other health centres—is pointed out (Figure 3).

Obviously, the choice of the criteria affects the results of site search in a substantial way.

Such an analysis is based on the verification of 13 combinations of weights vectors (W^{opt}_i), representative of all possible logically coherent combinations of weights.

The results show that the solution seems to be more sensitive to the passive mobility criterion than to other criteria. Varying the relative importance of proximity and per capita income criteria does not usually affect the results in a sensitive way. As for the indifference area criterion, it seems to affect the results, but only when its weight is high. Thus, we can assume that for reasonable combinations of the weights vectors the passive mobility criterion is the only one affecting the solution.

Figure 2 . Weighted overlay results

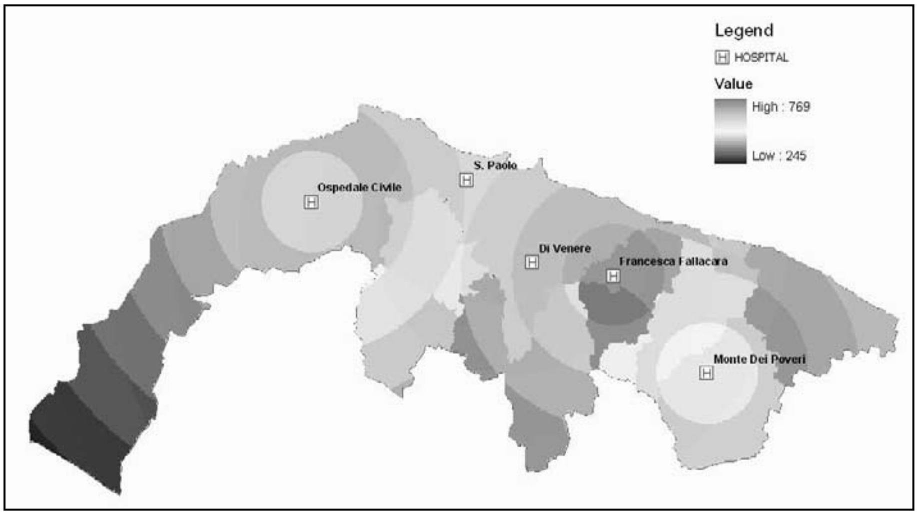
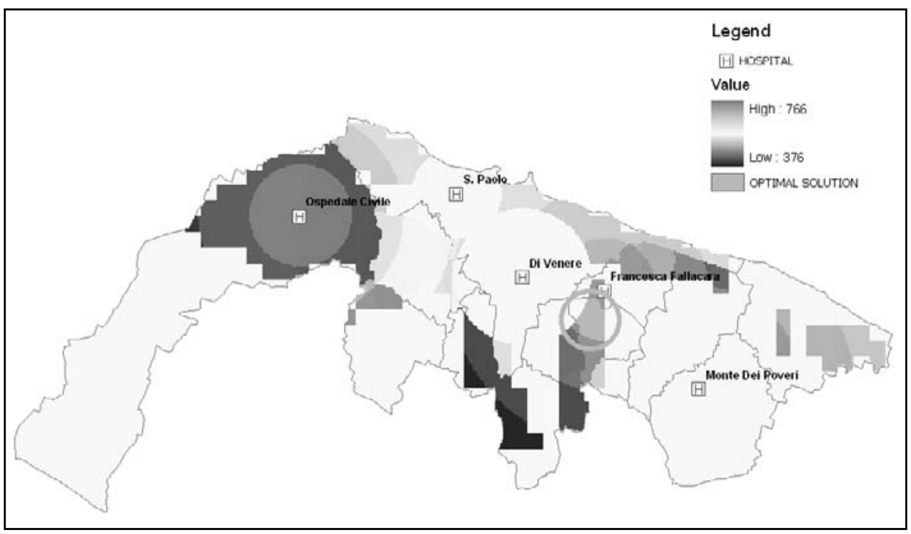


Figure 3 . Optimal solution using W^{opt}



At this stage, the procedure has to be extended to the localization of surgical and diagnostic services. Then, a suitability analysis for adopted criteria is carried out to check whether it is necessary to formulate some of them differently for surgical and diagnostic services.

Regarding to the diagnostic services, the rate of potential patients may be underestimated, because a part of the population may require diagnostic services not related to principal diseases. The per capita income criterion could lose importance with regard to healthcare services which do not provide essential assistance (e.g., plastic surgery).

As a rule, the criteria exclusively based on localization features and/or socio-demographic attributes associated with the healthcare district area, can be considered as independent from the specific healthcare service, and consequently, have general validity. Vice versa, for the criteria which adopt the point of view of the healthcare service provider (local government), it is useful to set the analysis and characteristic parameters according to each case.

Table 1. Sensitivity analysis

In Table 1, the results of the sensitivity analysis conducted to verify the variability of the optimal solution according to the variation of criteria relative weights are shown. W^{opt}_i	Different optimal solutions
W_1= [0.25, 0.25, 0.25, 0.25]	Yes
W_2= [0.55, 0.15, 0.15, 0.15]	Yes
W_3= [0.70, 0.10, 0.10, 0.10]	Yes
W_4= [0.10, 0.10, 0.10, 0.70]	Yes
W_5= [0.07, 0.07, 0.28, 0.58]	No
W_6= [0.03, 0.17, 0.40, 0.40]	No
W_7= [0.03, 0.40, 0.17, 0.40]	No
W_8= [0.04, 0.21, 0.21, 0.54]	No
W_9= [0.07, 0.58, 0.07, 0.28]	No
W_{10}= [0.15, 0.55, 0.15, 0.15]	No
W_{11}= [0.07, 0.58, 0.28, 0.07]	No
W_{12}= [0.04, 0.21, 0.54, 0.21]	No
W_{13}=[0.15, 0.15, 0.55, 0.15]	No

In particular, the geographical proximity criterion originates from the preference for opening new services in existing structures, to reduce the opening cost, depending on infrastructure and facilities costs. In the specific case of surgical or diagnostic services, technology, and accessibility components must be added to the cost function on which is based the proximity criterion. Consequently, it is necessary to implement a much more complex *opportunity/cost function* (C_{opp}), having the following expression:

$$C_{opp}= C_{acc}+C_{op}+C_{tec}$$

where:

C_{acc} = cost which patients have to pay to reach the surgical/diagnostic services;

C_{op} = cost of new surgical/diagnostic services, in term of infrastructures and facilities;

C_{tec} = cost of new medical technology needed.

According to all parameters of surgical service, it is assumed a rate of 2‰ for potential patients and the maps of suitable areas, corresponding to each criterion, are created.

The adoption of the new opportunity/cost function leads to create the new map of suitable areas by overlaying three independent maps, having the same relative importance, each one representing one component of the C_{opp}. In detail, the opening cost (C_{op}) and the accessibility cost (C_{acc}) are assumed to be rising proportionally with the euclidean distance from existing healthcare centres, with the rising ratio of the C_{acc} higher than the C_{op}. The technological cost is considered independent from distance.

As in the previous case, the weight vector used is W_{opt} = [0.03; 0.32; 0.21; 0.44].

The obtained solution is the same as the basic healthcare services (medicine). This demonstrates a quite high level of consistency for the adopted methodology.

With reference to this, it is conducted a new sensitivity analysis as in the first case, with the result that more combinations of weights vectors (W^{opt}_i) compared with the previous case, provided the optimal solution. Then the introduction of the opportunity/cost function seems to increase the effectiveness of the methodology.

FUTURE TRENDS

The previously mentioned trend towards a reorganisation of healthcare services, according to integration criteria, needs the implementation of an adequate system able to monitor network capabilities. GISs will play a key role in the definition of a planning process of healthcare districts able to manage resources in an efficient way and to determine a more coherent distribution of services, by this matching more adequately the population needs.

CONCLUSION

This article proposes an innovative methodology for finding a solution to the problem of localization of healthcare services in a healthcare network. Such a problem has been solved by applying principles and techniques, which are typical of the spatial analysis for localization problems, characterized by quantitative and qualitative, geographical and demographical factors.

The analysis is conducted using GIS technologies, which have great potentialities in terms of management and analysis of a large amount of heterogeneous data. Because of such features, GIS technologies are used as a decision support tool, by means of which the most suitable place for opening a new healthcare service in a healthcare district area has been identified. Such a methodology proves to be robust enough, with respect to the specific context considered, and solutions take into consideration all the main variables of the problem.

REFERENCES

Benabdallah, S., Wright, J.R. (1992). Multiple subregion allocation models. *Journal of Urban Planning and Development, ASCE, 118*(1), 24–40.

Brown S. (1998). Prenatal care: Reaching mothers reaching children. Washington, DC: Institute of Medicine, National Academies Press.

Carver, S.J. (1991). Integrating multi-criteria evaluation with geographical information systems. *International Journal of Geographical Information Systems, 5*(3), 321–339.

Chan, Y. (2001). *Location theory and decision analysis*. South Western College Publishing.

Collins M.G., Steiner F.R., & Rushman M.J. (2001). Land-use suitability analysis in the United States: Historical development and promising technological achievements. *Environmental Management, 28*(5), 611-621.

Cromley, R.G., & Hanink, D.M. (1999). Coupling land use allocation models with raster GIS. *Journal of Geographical Systems, 1*(2), 137–153.

Curtis, S., & Taket, A. (1996). *Health and societies: Changing perspectives*. London: Arnold.

Digregorio, G. (2006). *The localization of health services: Decision support systems based on hub and spoke model*. Unpublished doctoral thesis, Politecnico di Bari.

Eastman, J.R., Kyem, P.A.K., & Toledano, J. (1993). *GIS and decision making*. Genève: UNITAR.

Fortney, J., Rost, K., & Warren, J. (2003). Comparing alternative methods of measuring geographic access to health services. *Health Services and Outcomes Research Methodology, 1*(2), 173–184.

Harper, P.R., Shahani, A.K., Gallagher, J.E., & Bowie, C. (2004). Planning health services with explicit geographical considerations: A stochastic location–allocation approach. *The International Journal of Management Science, 33*, 141–152.

Hopkins L. (1977). Methods for generating land suitability maps: A comparative evaluation. *Journal of American Institute of Planners, 34*(1), 19-29.

Lin, H.Y., & Kao, J.J. (1998). A vector-based spatial model for landfill siting. *Journal of Hazardous Materials, 58*, 3–14.

Malczewski, J. (2004). GIS-based land-use suitability analysis: a critical overview. *Progress in Planning, 62*(1), 3–65.

Mullner, R.M., Chung, K., Croke, K.G., & Mensah, E.K. (2004). Geographic information systems in public health and medicine. *Journal of Medical Systems, 28*(3), 215–221.

Saaty, T.L. (1990). How to make a decision: The analytic hierarchy process. *European Journal of Operational Research, 48*(1), 9–26.

KEY TERMS

Geographical Information System (GIS): An information system able to collect both geographical and alphanumerical data.

Location Analysis: A decision-making procedure aimed at identifying the most suitable location within a set (site selection) or without a pre-established set (site search) for a specific purpose.

Multicriteria Analysis: A decision-making approach aimed at optimising different functions/variables at the same time.

Optimisation Problem: A problem consisting in either maximizing or minimizing one or more functions.

Simple Additive Weighting: A multicriteria decision-making technique consisting in assigning to each alternative a sum of values, each one associated to the corresponding evaluation criterion, and weighted according to the relative importance of the corresponding criterion.

Decision Support Systems for Cardiovascular Diseases Based on Data Mining and Fuzzy Modelling

Markos G. Tsipouras
University of Ioannina, Greece

Themis P. Exarchos
University of Ioannina, Greece

Dimitrios I. Fotiadis
University of Ioannina, Greece
Michaelideion Cardiology Center, Greece
Biomedical Research Institute, Greece

Aris Bechlioulis
University of Ioannina, Greece
Michaelideion Cardiology Center, Greece

Katerina K. Naka
University of Ioannina, Greece
Michaelideion Cardiology Center, Greece

INTRODUCTION

The widespread availability of new computational methods and tools for data analysis and predictive modelling requires medical informatics researchers and practitioners to systematically select the most appropriate strategy to cope with clinical prediction problems. In particular, data mining techniques offer methodological and technical solutions to deal with the analysis of medical data and construction of decision support systems. Furthermore, fuzzy modelling deals with the ambiguity inherent in all medical problems. These methods can be used to design and develop clinical decision support systems (CDSSs), which, after evaluated from the experts, can be integrated into clinical environments.

Cardiovascular diseases (CVDs) are the leading cause of death in many countries worldwide. According to the World Health Organization, CVDs are the cause of death of 16.6 million people around the globe each year. The multifaceted nature of these diseases, combined with a wide variety of treatments and outcomes, and complex relationships with other diseases, for example, diabetes, have made diagnosis and optimal treatment of cardiovascular diseases a problem for all but experienced cardiologists.

This article addresses the decision support regarding cardiovascular diseases, using computer-based methods, focusing on the coronary artery disease (CAD) diagnosis and on the prediction of clinical restenosis in patients undergoing angioplasty. Methods reported in the literature are reviewed with respect to (i) the medical information that are employing in order to reach the diagnosis and (ii) the data analysis techniques used for the creation of the CDSSs. In what concerns medical information, easily and noninvasively-obtained data present several advantages compared to other types of data, while data analysis techniques that are characterized by transparency regarding their decisions are more suitable for medical decision making. A recently developed approach that complies with the above requirements is presented. The approach is based on data mining and fuzzy modelling. Using this approach, one CDSS has been developed for each of the two cardiovascular problems mentioned above. These CDSSs are extensively evaluated and comments about the discovered knowledge are provided by medical experts. The later is of great importance in designing and evaluating CDSSs, since it allows them to be integrated into real clinical environments.

BACKGROUND

Data Mining is the process of discovering patterns and correlations from large amounts of data, using artificial intelligence, statistical, and mathematical techniques (Tan, Steinbach, & Kumar, 2005). Fuzzy logic is the extension of the classical crisp (binary) logic into a multivariate form. Fuzzy logic is closer to the human logic, thus being able to deal with real world noisy and imprecise data (Wang, 1986). CDSSs are computerized tools developed to assist physicians through the process of decision making. A known approach for the development of CDSSs is the use of experts' knowledge combined with an inference engine. However, recent advances in designing CDSSs employ automated knowledge extraction from data, using data mining techniques, while fuzzy logic provides several advantages in designing inference engines, compared to the classical crisp logic. The combination of data mining and fuzzy modelling provides a powerful tool for fully automated creation of CDSSs, experiencing several advantages: (i) transparency in decision making, (ii) addressing the ambiguity inherent in clinical data, and (iii) ability to interpret all decisions in a medical manner. All the above are of great importance for physicians, when performing decision making.

Coronary artery disease (CAD) is the development of atherosclerotic plaques in coronary arteries, resulting in coronary luminal narrowing and subsequently occlusion, and thus leading to myocardial infarction or sudden cardiac death. Coronary angiography (CA) is considered to be the "gold standard" method for the diagnosis of CAD and it is widely used. However, CA is an invasive and costly procedure that needs high level technical experience and technology and cannot be used for screening of large populations or close follow-up of treatment (Escolar, Weigold, Fuisz, & Weissman, 2006). Computer aided methodologies for CAD diagnosis have also been proposed in the literature; in this case the data obtained by some of the above mentioned methods or other sources (i.e., laboratory examinations, demographic, and/or history data, etc.) are evaluated by a computer-based application, leading to a CAD diagnosis. These methodologies can be divided into various categories, based on the type of data they use for subject characterization. Several methods are based on heart sounds associated with coronary occlusions (Akay & Welkowitz, 1993). Also, methods which employ the resting or exercise electrocardiogram (ECG) of the patient, extracting features from it, like the R wave (Szildgyi, Szildgyi, & David, 1997), the QT interval (Ng, Wong, Mora, Passariello, & Almeida, 1998), the T wave amplitude (Sabry-Rizk et al., 1999), the heart rate variability (Tkacz & Kostka, 2000), and the ST segment (Lewenstein, 2001) have been proposed. Furthermore, there are methods using medical images, such as SPECT (Haddad, Adlassnig, & Porenta, 1997), and methods based on arterio-scillography (Pouladian, Golpayegani, Tehrani-Fard, & Bubvay-Nejad, 2005). There exist also methods that employ demographic, history, and laboratory data (Frossyniotis et al., 2001; Mobley, Schechter, Moore, McKee, & Eichner, 2005; Tsipouras et al., 2006) and methods that combine more than one type of data (Kukar, Kononenko, Groselj, Kralj, & Fettich, 1999; Scott, Aziz, Yasuda, & Gewirtz, 2004).

The evolution and widespread adoption of percutaneous transluminal coronary angioplasty (PTCA) represents a major advance in the management of acute coronary syndromes, resulting in a significant reduction in early and late mortality compared to pharmacologic reperfusion therapy. Coronary artery restenosis remains a major limitation of PTCA and is usually defined as ≥ 50% stenosis in the treated segment at follow up, or at least 50% loss of the original gain in the minimal luminal diameter. Many clinical, angiographic, and procedural features have been studied as predictors for restenosis but it has proven difficult to stratify patients with regard to the risk of restenosis. Knowledge of risk factors for restenosis may help to refine indications of PTCA, reduce the frequency of restenosis, and select optimal candidates for a PTCA procedure. Despite lowering the restenosis rate with the implantation of coronary stents, it occurs in approximately 12-60% of the patients within 6 months after intervention depending mainly on the patients' and procedural characteristics. Computer aided methodologies for the prediction of clinical restenosis have also been proposed in the literature. These can be divided into two categories, regarding the data they use for the analysis: methods that employ only CAD risk factors such as demographic, history, and clinical data (Budde, 1999; Tsipouras et al., 2006) and methods that combine CAD risk factors with angiographic features (Maier, Mini, Antoni, Wischnewski, & Meier, 2001; Resnic, Popma, Ohno-Machado, 2000).

CDDSS FOR CARDIOVASCULAR DISEASES

A recently developed method for automated CDSSs creation is based on data mining and fuzzy modeling (Tsipouras et al., 2006). Specifically, in order to create the CDSS, a three stage methodology is used: (i) creation of a rule-based classifier using data mining techniques, (ii) development of a fuzzy logic model, and (iii) optimisation of the fuzzy logic model's parameters. Briefly, in the first stage, a set of crisp rules is generated. This is performed by inducing a decision tree from a training dataset and then transforming the tree into a set of rules. In the second stage, the crisp set of rules is transformed to a fuzzy set of rules, using a membership function instead of the crisp ones and S and T norms instead of the binary AND and OR operators. Finally, in the third stage, all thresholds and parameters involved in the fuzzy model are optimized with respect to a training dataset. The fuzzy model with the optimized parameters comprises the final CDSS. To apply the above methodology in a specific domain, an annotated dataset is required. The quality of this dataset is very important for developing an effective CDSS. This methodology has been used in order to create CDSSs for CAD diagnosis and prediction of clinical restenosis in patients undergoing angioplasty.

Application to CAD Diagnosis

The dataset used for CAD diagnosis included 199 subjects suspected of having CAD and who underwent coronary angiography for the first time. Patients with known CAD were excluded from the study. Of the subjects, 89 had normal angiograms and in the remaining 110 subjects the presence of CAD was confirmed by two experts. In order to characterise each subject, 19 features (shown in Table 1) were used. Two demographic features were recorded: the age and sex of the patient. From the subject's history data, the family history of CAD, smoking history, history of diabetes mellitus, and measurements of hypertension or hyperlipidaemia were used. The incorporated laboratory investigations were creatinine, glucose, total cholesterol, high density lipoprotein, and triglycerides. In addition, Carotid-Femoral Pulse Wave Velocity and Augmentation index were also used as non-invasive indices of arterial stiffness (Van Bortel et al., 2002; Woodman & Watts, 2003). In order to confirm the presence or absence of CAD, coronary angiography was performed by the Judkins technique. All coronary angiograms were visually assessed by two experienced cardiologists to reach consensus agreement. Significant CAD was defined as at least one stenosis of 50% or greater diameter in at least one coronary artery vessel. The absence of CAD was defined as completely smooth epicardial coronary arteries.

In Figure 1 an indicative set of rules is presented. These rules are extracted using data mining and (before fuzzification) were commented by the experts. According to the experts, gender is the most important feature in the produced set of rules. However, this is partially driven from the dataset since 64% of our male population (98/152) was diagnosed with CAD as compared with 25% (12/47) in the female population. Smoking has also been proven to be an important marker for CAD prediction in women; 77% of non smoking women did not suffer from CAD (33/42). In the male population, low HR (i.e., <=49 beats per minute) was found to be an important predictor of CAD; this might be explained by the use of β-blockers (antianginal medications that lower HR) in subjects with very high clinical suspicion of CAD. Also, it appears that elderly males (i.e., age>69 years) with symptoms and signs of CAD have high probability to be diagnosed with CAD, since CAD was found in 92% (24/26) of our elderly male population. In males aged less than 69 years, family history of CAD appears to be a relatively important diagnostic feature for CAD (76% of those with positive family history had CAD, i.e., 26/34). However, some of the derived rules cannot be fully explained based on standard medical knowledge, mainly due to the data-driven nature of the proposed method, which can also discover non important and spurious rules.

In Table 2, several computer aided diagnosis methodologies for CAD are presented. Some of the non-invasive methods, such as computerized tomography or magnetic resonance imaging, suffer from similar problems as CA, that is, being costly and requiring specialized technology and expertise, while they are not widely available (Escolar et al., 2006). Most of the computer based methods are based on the analysis of data obtained by examinations, such as stress ECHO and SPECT, which are also expensive, not widely available, and suffer from technical limitations (Merz, 2005). Exercise stress testing is inexpensive and widely available, but cannot be applied to all patients and has low sensitivity and specificity in the diagnosis of CAD.

Table 1. Features for CAD CDSS

#	Feature	Units
1	Age	Years
2	Sex	male(1), female(0)
3	Family History (FH)	yes(1), no(0)
4	Smoking (Sm)	smoker (2), ex-smoker (1), non-smoker (0)
5	Diabetes mellitus (DM)	FBGC $\geq$ 126mg/dl (1) else (0)
6	Hypertension (HT)	DBP>90mmHg and/or SBP>140mmHg (1) else (0)
7	Hyperlipidemia (HL)	total cholesterol over 220mg/dl (1) else (0)
8	Creatinine (Cre)	mg/dL
9	Glucose (Glu)	mg/dL
10	Total Cholesterol (TC)	mg/dL
11	High Density Lipoprotein (HDL)	mg/dL
12	Triglyceride (TG)	mg/dL
13	Body Mass Index (BMI)	kg/ m^2
14	Waist	Cm
15	Heart Rate (HR)	Bpm
16	Systolic Blood Pressure (SBP)	mmHg
17	Diastolic Blood Pressure (DBP)	mmHg
18	Carotid femoral pulse wave velocity (PWVcf)	m/sec
19	Augmentation Index (AIx)	%

The populations and the parameters collected and analyzed differ among studies; in some of the studies, only male subjects (Lewenstein, 2001; Mobley et al., 2005) or subjects with previous myocardial infarction (MI) or coronary artery bypass grafting (CABG) (Kukar et al., 1999), were included. It should be mentioned that most of the methods reported in Table 2 are based on neural networks (Akay & Welkowitz, 1993, Frossyniotis et al., 2001; Kukar et al., 1999; Mobley et al., 2005; Scott et al., 2004). These methods are not able to provide clear interpretation for their decisions.

Application to Prediction of Clinical Restenosis in Patient Undergoing Angioplasty

The dataset used for prediction of clinical restenosis in patients undergoing angioplasty consisted of 1,000 subjects that underwent angioplasty. In order to characterise the subjects, the 15 features shown in Table 3 were used. Family history, hypertension, diabetes mellitus, current smoking, and hyperlipidemia were defined as for the CAD CDSS. Clinical presentation of CAD was classified as unstable angina, stable angina,

Table 2. Comparison of several computer aided methodologies for CAD diagnosis

Author – Year	Number of subjects	Method – analysis	Se (%)	Sp (%)	Acc (%)
Akay & Welkowitz, 1993	112	Heart sounds – Neural network	78	89	
Haddad et al., 1997	100	SPECT – Case based reasoning	98	70	93
Kukar et al., 1999	327	Subject's data, exercise ECG, SPECT – Neural network	96	84	92
Frossyniotis et al., 2001	139	Exercise ECG, subject's data, indices of arterial stiffness – Neural network	78	75	78
Lewenstein, 2001	776	Exercise ECG	97	98	
Scott et al., 2004	102	SPECT, subject's data – Neural network	88	65	
Mobley et al., 2005	2004	Subject's data – Neural network	100	26	
Pouladian et al., 2005	51	Arterio-oscillography – signal processing	73	90	
Tsipouras et al., 2006	199	Subject's data, indices of arterial stiffness – data mining, fuzzy modelling	80	65	73

or acute myocardial infarction. The vessels treated with angioplasty were the right coronary artery, left main stem, left anterior descending artery, left circumflex artery, and bypass grafts. The patients underwent either angioplasty with balloon alone or balloon followed by stenting with a noncoated metal stent. All patients were followed up for at least 12 months. The composite end point of the study was clinical restenosis manifested as cardiac death, a new non fatal myocardial infarction or a new revascularisation attempt of the stented vessel in less than 6 months after the initial angioplasty procedure.

In Figure 2, some indicative rules are presented. The experts agree that the most important feature in the production of rules for the prediction of clinical restenosis is the number of diseased coronary vessels since 15.4% (i.e., 87/565) of the multivessel cases presented with restenosis in less than 6 months after the percutaneous coronary intervention (PCI) procedure in contrast to only 7.9% (i.e., 29/365) of the single vessel patients. Patients who underwent the angioplasty in a stable condition (i.e., stable angina) had probably a favourable prognosis (only 8.3% of patients with stable angina presented with restenosis in less than 6 months compared to 13.5% of those with unstable coronary syndromes), except perhaps for older (i.e., > 65 years old) or diabetic people. On the other hand, unstable coronary syndromes (i.e., unstable angina and myocardial infarction) are related to worse prognosis in terms of clinical restenosis even in younger ages (i.e., 55 years old) or patients without many severe cardiovascular risk factors. History of a previous PCI procedure and especially history of coronary aortic bypass surgery (occurrence of restenosis in 19.4% of patients with previous CABG compared to 12.2% without CABG history) are associated with increased risk of restenosis after coronary angioplasty. However, some of the derived rules could not be explained based on current medical knowledge on various interactions among the features used in our models.

In Table 4, a comparison of several computer-aided methodologies for the prediction of clinical restenosis is presented. Different datasets were used in each method; therefore a direct comparison is not feasible. Some approaches (Budde, 1999; Tsipouras et al., 2006) are based on noninvasively acquired data, while others employ also data obtained from angiographies (Maier et al., 2001; Resnic et al., 2000), thus being invasive approaches. Data analysis is performed mainly using rule based systems (Budde, 1999, Maier et al., 2001,

Figure 1. Indicative rules (crisp) for CAD diagnosis

IF (Sex = 0 and Sm = 0 and HR ≤ 65 and PWVcf ≤ 10.5 and AIx ≤ 48 and FH = 0)	THEN	normal
IF (Sex = 0 and Sm = 0 and HR ≤ 65 and PWVcf ≤ 10.5 and AIx ≤ 48 and FH = 1 and TC ≤ 240)	THEN	normal
IF (Sex = 0 and Sm = 0 and HR ≤ 65 and PWVcf ≤ 10.5 and AIx ≤ 48 and FH = 1 and TC > 240)	THEN	CAD
IF (Sex = 0 and Sm = 0 and HR ≤ 65 and PWVcf ≤ 10.5 and AIx > 48)	THEN	CAD
IF (Sex = 0 and Sm = 0 and HR ≤ 65 and PWVcf > 10.5)	THEN	CAD
IF (Sex = 0 and Sm = 0 and HR > 65)	THEN	normal
IF (Sex = 0 and Sm = 2 and Glu ≤ 94)	THEN	CAD
IF (Sex = 0 and Sm = 2 and Glu > 94)	THEN	normal
IF (Sex = 1 and HR ≤ 49)	THEN	CAD
IF (Sex = 1 and HR > 49 and Age ≤ 69 and FH = 0 and DM = 0 and BMI ≤ 27.99 and TRG ≤ 191 and Age ≤ 50)	THEN	normal
IF (Sex = 1 and HR > 49 and Age ≤ 69 and FH = 0 and DM = 0 and BMI > 27.99 and HR > 53 and HDL > 27)	THEN	normal
IF (Sex = 1 and HR > 49 and Age ≤ 69 and FH = 1 and HT = 0)	THEN	CAD
IF (Sex = 1 and HR > 49 and Age ≤ 69 and FH = 1 and HT = 1 and Glu ≤ 103)	THEN	normal
IF (Sex = 1 and HR > 49 and Age ≤ 69 and FH = 1 and HT = 1 and Glu > 103)	THEN	CAD
IF (Sex = 1 and HR > 49 and Age > 69 and DM = 0)	THEN	CAD
IF (Sex = 1 and HR > 49 and Age > 69 and DM = 1 and Cre ≤ 1)	THEN	normal
IF (Sex = 1 and HR > 49 and Age > 69 and DM = 1 and Cre > 1)	THEN	CAD

Figure 2. Indicative rules (crisp) for prediction of clinical restenosis

IF (SVD = 1 and CP = 1 and VT = 1 and FH = 0 and HL = 1 and PTCA = 0)	THEN	normal
IF (SVD = 1 and CP = 1 and VT = 1 and FH = 0 and HL = 1 and PTCA = 1)	THEN	restenosis
IF (SVD = 1 and CP = 1 and VT = 1 and FH = 1 and ST = 1)	THEN	normal
IF (SVD = 1 and CP = 1 and VT = 2)	THEN	normal
IF (SVD = 1 and CP = 1 and VT = 3 and Age ≤ 66)	THEN	normal
IF (SVD = 1 and CP = 2 and CABG = 0 and HL = 0 and HT = 0)	THEN	normal
IF (SVD = 1 and CP = 2 and CABG = 1)	THEN	restenosis
IF (SVD = 1 and CP = 3)	THEN	normal
IF (SVD = 0 and CP = 1 and PTCA = 0 and VT = 1 and CABG = 0 and Age ≤ 51)	THEN	normal
IF (SVD = 0 and CP = 1 and PTCA = 0 and VT = 2 and Sm = 0 and HT = 0 and Age > 60)	THEN	restenosis
IF (SVD = 0 and CP = 1 and PTCA = 0 and VT = 3 and ST = 1 and FH = 0 and Age ≤ 58)	THEN	restenosis
IF (SVD = 0 and CP = 2 and Sex = 0 and VT = 3 and DM = 0 and ST = 2)	THEN	restenosis
IF (SVD = 0 and CP = 2 and Sex = 0 and VT = 3 and DM = 1)	THEN	restenosis
IF (SVD = 0 and CP = 2 and Sex = 1 and PTCA = 1)	THEN	restenosis
IF (SVD = 0 and CP = 3 and HT = 0 and FH = 0 and ST = 1 and Age ≤ 66)	THEN	normal
IF (SVD = 0 and CP = 3 and HT = 0 and FH = 0 and ST = 1 and Age > 66)	THEN	restenosis
IF (SVD = 0 and CP = 3 and HT = 1 and Sm = 0 and PTCA = 0 and Sex = 1 and DM = 0)	THEN	normal
IF (SVD = 0 and CP = 3 and HT = 1 and Sm = 0 and PTCA = 0 and Sex = 1 and DM = 1)	THEN	restenosis

Table 3. Features for prediction of clinical restenosis CDSS

#	Feature	Units
1	Stent Type (ST)	Balloon (0), Bare Metal Stent (1), Drug Eluting Stent (2)
2	Sex	male(1), female(0)
3	Age	Years
4	Diabetes Melitus (DM)	FBGC ≥ 126mg/dl (1) Else (0)
5	Hypertension (HT)	DBP>90mmHg and/or SBP>140mmHg (1) else (0)
6	Smoking (Sm)	smoker (1), non-smoker (0)
7	Hyperlipidemia (HL)	total cholesterol over 220mg/dl (1) else (0)
8	Family History (FH)	yes(1), no(0)
9	CAD History (CAD)	yes(1), no(0)
10	Prior PTCA (PTCA)	yes(1), no(0)
11	Prior CABG (CABG)	yes(1), no(0)
12	Single Vessel Disease (SVD)	yes(1), no(0)
13	Clinical Presentation (CP)	Unstable angina (1), Acute myocardial infarction (2), Stable angina (3)
14	Vessel Treated (VT)	Left anterior descending (1), Left circumflex (2), Right coronary artery (3), Left main (4), Bypass graft (5)
15	IIB/IIIA	yes(1), no(0)

Table 4. Comparison of several computer aided methodologies for prediction of restenosis after coronary angioplasty

Author - Year	Number of subjects	Data	Method	Accuracy (%)
Budde, 1999	2500	Risk factors for CAD	Rule based system	95
Resnic et al., 2000	2804	Demographic, clinical and angiographic data	Statistical analysis Risc score ANN	81 79 81
Maier et al., 2001	325 (lesions)	Clinical and angiographic data	Statistical analysis Rule based system	58 92
Tsipouras et al., 2006	1000	Demographic, history and clinical data	Data mining, fuzzy modelling	61

Tsipouras et al., 2006) or artificial neural networks (Resnic et al., 2000).

FUTURE TRENDS

Researchers have spent great efforts in the design and development of CDSSs for several domains in medicine and health care. Most of the CDSSs try to reduce the effort and time of the experts when performing diagnosis. However, a plethora of systems has been presented for staging, treatment, dose adjustment, and follow-up. An important requirement of this type of system is the transparency regarding the automated decisions and the interpretation they provide. Moreover, experts rely more on systems that their automated decisions coincides with established medical knowledge. Integration of knowledge provided by the experts and knowledge generated using data mining methods, for the creation of more sophisticated CDSSs, is the trend of the future. Established medical knowledge combined with data mining models is the key to increase the effectiveness of these systems and provide advanced tools for computer based clinical medicine.

CONCLUSION

CVDs are among the most life threatening diseases worldwide. A vast amount of patients suffering from CVDs are examined, hospitalized, and treaded every day; this has a major impact in national health care systems. The importance of CVDs has lead to the development of a large number of computer-based CDSSs that mainly focus in diagnosis, treatment, or follow-up. An important requirement of a CDSS is its ability to provide transparency regarding the generated decisions, thus providing a clear insight of their inner process for decision making; this is essential for physicians in order to incorporate such systems in their clinical practice. Another important requisite is an extensive evaluation of a CDSS by medical experts, in order to comment on the functionality of the system and fully exploit its potential. Furthermore, the employment of easily obtained and noninvasive features for decision support is considered an advantage since it facilitates the straightforward application of the system. Most of the CDSSs proposed in the literature do not meet these requirements, thus complicating their application in clinical practice. The CDSSs presented in this article fulfill these issues, making them suitable to be used in clinical practice.

REFERENCES

Akay, M., & Welkowitz, W. (1993). Acoustical detection of coronary occlusions using neural networks. *Journal of Biomedical Engineering, 15,* 469-473.

Budde, T., Haude, M., Hopp, H. W., Kerber, S., Caspar, G., Fa, G., Fingerhut, M. et al. (1997). A prognostic computer model to predict individual outcome in interventional cardiology: The INTERVENT project. *European Heart Journal, 18,* 1611-1619.

Escolar, E., Weigold, G., Fuisz, A., & Weissman, N.J. (2006). New imaging techniques for diagnosing coronary artery disease. *Canadian Medical Association Journal, 174*(4), 487-495.

Frossyniotis, D., Vrettos, S., Vogiatzis, D., Stafulopatis, A., Konstantinopoulos, C., Fotiadis, D.I. et al. (2001). An intelligent system for the early diagnosis of coronary artery disease. In *Proceedings of Neural Networks and Expert Systems in Medicine and Healthcare*.

Haddad, M., Adlassnig, K.P., & Porenta, G. (1997). Feasibility analysis of a case-based reasoning system for automated detection of coronary heart disease from myocardial scintigrams. *Artificial Intelligence in Medicine, 9,* 61-78.

Kukar, M., Kononenko, I., Groselj, C., Kralj, K., & Fettich, J. (1999). Analysing and improving the diagnosis of ischaemic heart disease with machine learning. *Artificial Intelligence in Medicine, 16,* 25-50.

Lewenstein, K. (2001). Radial basis function neural network approach for the diagnosis of coronary artery disease based on the standard electrocardiogram exercise test. *Medical and Biological Engineering and Computing, 39,* 1-6.

Maier, W., Mini, O., Antoni, J., Wischnewski, M., & Meier, B. (2001). ABC stenosis morphology classification and outcome of coronary angioplasty: reassessment with computing techniques. *Circulation, 103*(9), 1225-1231.

Merz, N.B. (2005). Assessment of patients at intermediate cardiac risk. *American Journal of Cardiology, 96,* 2J-10J.

Mobley, B.A., Schechter, E., Moore, W.E., McKee, P.A., & Eichner, J.E. (2005). Neural network predictions of significant coronary artery stenosis in men. *Artificial Intelligence in Medicine, 34,* 151-161.

Ng, F., Wong, S., Mora, F., Passariello, G., & Almeida, D. (1998). Diagnostic value of QT interval hysteresis loop in coronary artery disease. In *Proceedings Computers in Cardiology* (pp. 417-420).

Pouladian, M., Golpayegani, M.R.H., Tehrani-Fard, A.A., & Bubvay-Nejad, M. (2005). Noninvasive detection of coronary artery disease by arterio-oscillography. *IEEE Transactions on Biomedical Engineering, 52,* 743-747.

Resnic, F.S., Popma, J.J. & Ohno-Machado, L. (2000). Development and evaluation of models to predict death and myocardial infarction following coronary angioplasty and stenting. In *Proceedings Americal Medical Informatics Association Symposium* (pp. 690-693).

Sabry-Rizk, M., El-Khafif, S., Carson, E., Zgallai, W., Grattan, K., Morgan C. et al. (1999). Suspicious polyphase patterns of normal looking ECGs provide fast early diagnoses of a coronary artery disease. In *Proceedings IEEE Engineering in Medicine and Biology* (Vol. 2, pp. 1391-1393).

Scott, J.A., Aziz, K., Yasuda, T., & Gewirtz, H. (2004). Integration of clinical and imaging data to predict the presence of coronary artery disease with the use of neural networks. *Coronary Artery Disease, 15*(7), 427-434.

Szildgyi, S.M., Szildgyi, L., & David, L. (1997). Assessment of R wave amplitude changes in coronary artery disease. In *Proceedings IEEE Engineering in Medicine and Biology* (Vol. 1, pp. 272-274).

Tan, P.N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Addison Wesley.

Tkacz, E.J., & Kostka, P. (2000). An application of wavelet neural network for classification of patients with coronary artery disease based on HRV analysis. In *Proceedings IEEE Engineering in Medicine and Biology* (Vol. 2, pp. 1391-1393).

Tsipouras, M.G., Exarchos, T.P., Papaloukas, C., Bechlioulis, A., Kotsia, A., Nanou, T. et al. (2006). Automatic creation of decision support systems: Application and results in the cardiovascular diseases domain. *The Journal on Information Technology in Healthcare, 4*(4), 222-230.

Van Bortel, L.M., Duprez, D., Starmans-Kool, M.J., Safar, M.E., Giannattasio, C., Cockroft, J. et al. (2002). Clinical applications of arterial stiffness, task force III: Recommendations for user procedures. *American Journal of Hypertension, 15*, 445-452.

Wang, L.X. (1986). *A course in fuzzy systems and control*. New York: Prentice-Hall.

Woodman, R.J., & Watts, G.F. (2003). Measurement and application of arterial stiffness in clinical research: Focus on new methodologies and diabetes mellitus. *Medical Science Monitor, 9,* RA101-109.

KEY TERMS

Artificial Neural Network: An interconnected group of artificial neurons that uses a mathematical or computational model for information processing based on a connectionist approach to computation.

Clinical Decision Support Systems (CDSS): Computer based methods that aim to assist clinicians in decision making. The core of CDSSs is an inference engine that can generate case specific advice based on medical data.

Clinical Restenosis: Death presumably from cardiac causes, myocardial infarction not attributable to another coronary artery than the target vessel, and target vessel revascularization either by repeat PTCA or CABG.

Coronary Artery Disease: The narrowing of the coronary arteries, sufficiently to prevent adequate blood supply to the heart muscle. It is usually caused by atherosclerosis and may progress to the damage of heart muscle.

Data Mining: The process of extracting previously unknown and potentially useful knowledge, hidden in large volumes of data.

Fuzzy Logic: A way of reasoning that can cope with uncertain or partial information.

Inference Engine: The part of a decision support system that performs the reasoning function.

Definition and Measurement of the Technology Readiness Concept

María Teresa Martín
Rey Juan Carlos University, Spain

María Victoria Román
Rey Juan Carlos University, Spain

Manuel Recio
University of Almería, Spain

INTRODUCTION

During the last few decades, various theoretical developments have been carried out with a view to describing the characteristic and distinct behavioral process that lies under any adoption of technological services and products. These developments are based mainly on the Social Psychology approach.

There are three extensive theories within the field of Social Psychology whose ultimate purpose has been to define the internal psychological factors that explain human behavior: the Expectancy-Value Theory, the Cognitive Dissonance Theory, and the Self-Perception Theory. While the Expectancy-Value Theory has been used widely in the research of adoption and usage of information systems, the other two theories have been less recognized.

Of all Expectancy-Value Theory models, we should draw our attention to the Reasoned Action Model (Azjen & Fishbein, 1980), because it underlies many of the studies on usage of technology. The Planned Behavior Model (Azjen, 1985, 1991) represents a reformulation of the Reasoned Action Model, justified by the existence of conducts that, albeit in part, a person cannot voluntarily keep under control. A rough description of both models is presented in this chapter, inasmuch as they served as a basis for the construction of the Technology Acceptance Model (Davis, 1989; Davis, Bagozzi & Warshaw, 1989), known as one of the main models for the technology readiness concept. The Technology Acceptance Model seems to possess a similar or even better explicating power than its predecessors (Davis et al., 1989; Mathieson, 1991; Taylor & Todd, 1995a; Chau & Hu, 2002).

Once we recognize the importance of attitude in predicting the use of technology, we continue by analyzing the Technology Readiness Model (Parasuraman & Colby, 1997, 1998, 2001). During the last few years, this model has been gaining more and more recognition among marketing professionals as a model for the explanation of the technology attitude.

Out of our review, we can make a synthesis of variables related to technology readiness by the analyzed models. The chapter ends with an exposition of the reasons that have led us to choose the Technology Readiness Model as a basic theoretical model to develop a scale of technology attitude.

Expectancy-Value Models

We now proceed with a brief examination of the main Expectancy-Value Models that have served as a theoretical basis for models of use of technology prediction, so we can gain a much better insight of the latter.

The Reasoned Action Model

The Reasoned Action Model (Azjen & Fishbein, 1980), whose ultimate aim is the prediction and understanding of the human behavior determinants, argues that the latter is mostly under the subject's control and, as a consequence, can be forecast by observing the declaration individuals make of their intentions to carry out or not to carry out a particular type of conduct. Also, the behavioral intention directly depends on two factors: a personal one or attitude toward behavior that represents the individual, either positive or negative, evaluation about whether an action should or should not be done; and a social factor or subjective norm that is the personal awareness of social pressures that force us to take that course of action.

Attitudes can be explained by considering the information the individuals possess that shapes their beliefs in the consequences when a certain conduct becomes real. Attitudes do not depend solely on beliefs, but also on the evaluation people make about each of those behavioral beliefs. Therefore, the drive of each belief (degree of certainty that upholds it) is boosted by the negative or positive intensity of the outcome of the conduct. The sum of all products results in an attitude.

Subjective norms can be assessed indirectly from two main components: first, the informative basis of individuals, which establishes normative beliefs, that is to say, beliefs about how other people or institutions (referees) judge how an individual should behave; second, the subject's motivation to observe the referees' directives. Like it happens with attitudes, each normative belief is boosted by the motivation to follow that opinion. The subjective norm is obtained from the sum of all these products.

In this model, we can see a display of another series of external variables or stimulus conditions that affect relationships between intentions and behavior by means of their direct influence on beliefs in the consequences of conduct and in normative beliefs (see Figure 2).

Azjen and Madden (1986) introduced some nuances into the Reasoned Action Model by pointing out that previous experience, attitude, and subjective norm can influence one another; attitudes can influence behavior both indirectly through behavioral intention and also directly; subjective norm can only have an indirect effect on behavior; and previous experience can influence conduct indirectly as well as directly. The results of this investigation also show that the relation between intention and behavior is not as intense as it should be to make trustworthy predictions. Thus, "intentions only forecast behaviors not requiring either special abilities or skills, extraordinary opportunities, nor the other people´s cooperation (Morales, Rebolloso & Moya, 1996, pp. 563-564).

The Planned Behavior Model

Azjen (1985, 1991) reformulated the aforementioned Reasoned Action Model, thus originating the Planned Behavior Model (see Figure 2). The difference between both models lies in the fact that the latter introduces the perceived behavioral control element to improve the behavioral intention forecast. The introduction of this intention-determining factor is justified by the existence of conducts that, at least in part, cannot be kept under the voluntary control of the individuals. That is to say, irrespective of a likely favorable attitude toward a certain behavior and the strong pressure that individuals receive around them urging them to do it, their intention will also influence the subjective likeliness of achieving it or perceived behavioral control.

Figure 2. Reasoned action model (Source: Adapted from Ajzen and Fishbein, 1980)

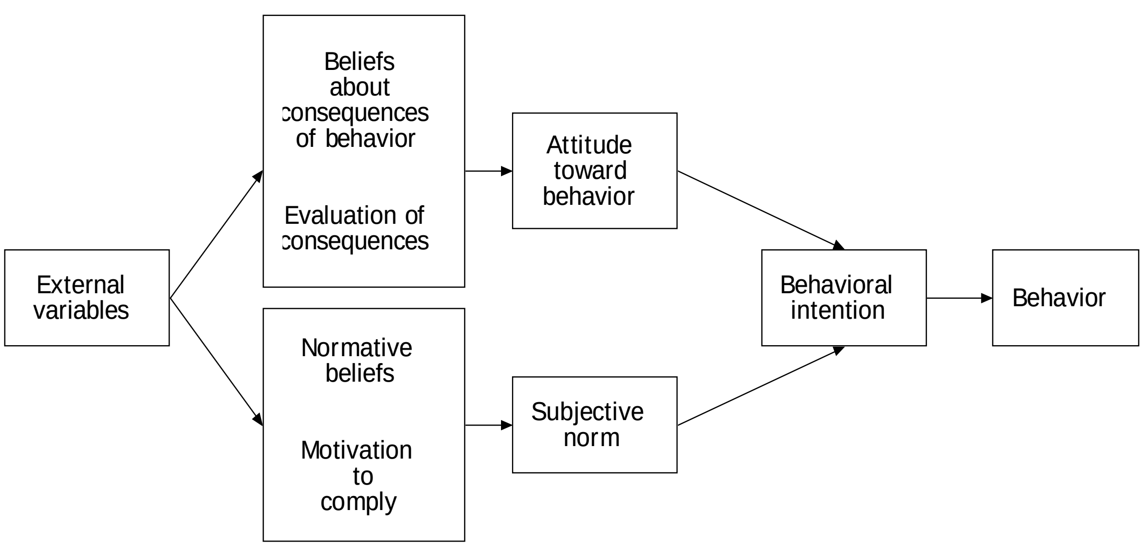
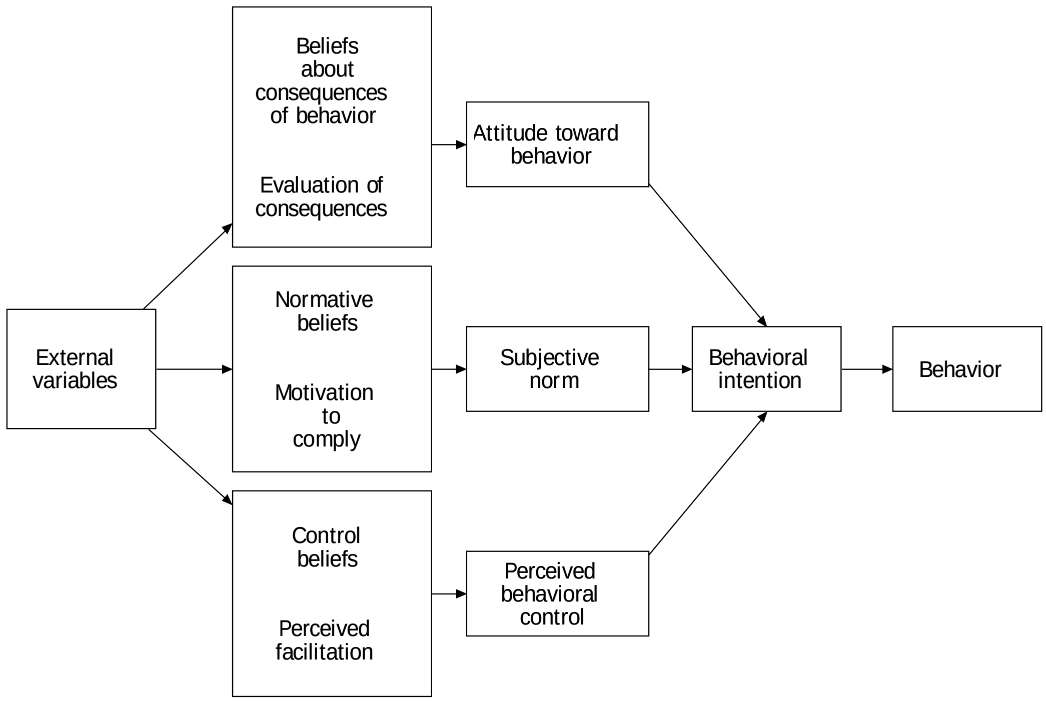


Figure 2. Planned behavior model (Source: Adapted from Ajzen, 1985, 1991)



Perceived behavioral control depends on control beliefs, which decide whether individuals possess or do not possess the necessary skills or resources to carry out a certain conduct, and the suitable opportunities. Each control belief is boosted by the enhancing or inhibitory effect of the resource or opportunity dealt with. The final outcome after summing up all these products is the perceived behavioral control.

As the perceived behavioral control nears the control effectively exerted by individuals on their respective conducts, the relationship with the latter is direct.

Prediction Models of Use of Technology

The use of technology represents a generally accepted measure of the success of its widespread presence. Therefore, the aim of many investigations has been to determine which factors influence this kind of behavior.

In this section, we will look into the main theoretical models that aim to describe the characteristic behavioral process underlying the adoption of technological services and products. Through their analysis, we can detect the importance given to technology attitude as a key factor in predictions of use of technology.

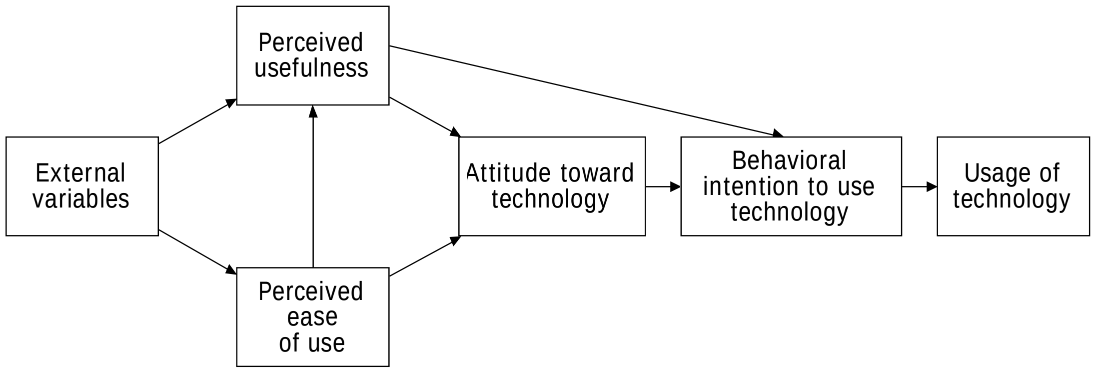
The review of the scientific literature has made clear that technology readiness is a multidimensional concept, although a consensus of opinion about the number and definition of these dimensions does not exist. In any case, such dimensions contain enhancing and inhibitory aspects of technology attitude.

With a view to completing the stage in which the domain of the studied concept is specified, after the exposition of each model, detailed aspects related to measure scales developed in various contexts and with different participants and technologies can be found. The examination of the various aspects of technology attitude (separately detected in the analyzed studies) will allow us to establish the suitability of its addition within the development of a measurement scale of the degree of technology readiness.

The Technology Acceptance Model

The Technology Acceptance Model (Davis, 1989; Davis et al., 1989) is one of the most referred-to models for the explanation of adoption of information technologies. The acceptance of technology is firmly defined by Gattiker (1984) as the psychological state of individuals with regard to their voluntary or desired use of a particular technology.

Figure 3. Technology acceptance model (Source: Adapted from Davis (1989) and Davis, et al. (1989))



Based on the Reasoned Action Model and the Planned Behavior Model, the Technology Acceptance Model put forward the idea that the behavior of adoption of a certain information system is determined by the intention of making use of it. In turn, such a behavioral intention is determined by technology attitude, which is influenced by two factors: on the one hand, beliefs in perceived usefulness, and on the other hand, beliefs in the perceived ease of use of the analyzed technology. Perceived usefulness can be defined as the extent to which individuals think using the information system will improve their working productivity. The perceived ease of use can be defined as the extent to which an individual thinks the use of a system will not require any effort. Although component perceived usefulness has a direct effect on the behavioral intention, the perceived ease of use influences such intention only indirectly.

Furthermore, both types of factors are exposed to the influence of external variables. By operating on these variables, control can be exerted on the beliefs of the information system users, thus indirectly influencing the intentions of use and the effective utilization of technology. The described model is shown in Figure 3.

Technology Attitude Scales with a Theoretical base on the Technology Acceptance Model

The Technology Acceptance Model has been put into practice by making use of different types of technologies as a social object of interchange and considering different user profiles as a sample (see Table 1).

A review of various studies allowed us to note that among their authors there is a widespread inclination to remove the attitude component out of the model. They consider that behavioral intention is determined by two main components: on the one hand, beliefs in perceived usefulness, and on the other hand, beliefs in the perceived ease of use of the analyzed technology. In that way, the perceived usefulness component has a direct effect on the behavioral intention. The perceived ease of use influences such intention, indirectly and directly, through the perceived usefulness component.

Other authors, although considering attitude as a component of the model, identified no significant relation between attitude and behavioral intention.

The results obtained cannot lead us to erroneous conclusions about the dilemma of consideration or non-consideration of attitude as a key factor when predicting the use of technology. In the reviewed investigations, when attitude is taken into account as a component of the model, it is categorized as a one-dimensional concept. That is why simple scales have been applied to measure it, following the ones suggested by Ajzen and Fishbein (1980) introduced in the Reasoned Action Model and Planned Behavior Model. From our point of view, an attitude scale constructed like this lacks validity of content insofar as it does not comprise all possible dimensions and contents of the analyzed concept.

The decision made by various authors to remove the attitude factor from the Technology Acceptance Model and, in its stead, to consider the direct effect of perceived usefulness and perceived ease of use factors on the behavioral intention, suggest to us that we should consider attitude as a multidimensional feature. Therefore, attitude is not something global. It is made up of a combination of a series of variables or dimensions that determine attitude toward technology.

Once we established which dimensions constitute the domain of the attitude concept and made clear that it is a multidimensional one, the priority of the following step was to identify items in accordance with the conceptual dimensions as found. Next, we displayed in Table 2 the items included in the measurement subscale of perceived usefulness in the analyzed studies, whereas in Table 3, we placed the items of the measurement subscale of perceived ease of use.

As we can see in the measurement of perceived usefulness, the use of the following items stands out:

- Using (information technology) increases my productivity.
- Using (information technology) increases my job performance.
- Using (information technology) enhances my effectiveness on the job.
- Overall, I find the (information technology) useful in my job.

Table 1. Application of technology acceptance model in various contexts

Software	User profile	Source
Text-editor	MBA students	Davis (1989)
Text-editor, e-mail	Managers and professionals	Davis (1993)
Spreadsheet	Students	Mathieson (1991)
Writeone, Chart-Master	MBA students	Davis et al. (1989)
Voicemail system, customer dial-up system	(Not specified)	Subramanian (1994)
University computing, resource center, business school student	Students	Taylor & Todd (1995a)
University computing, resource center, business school student	Students	Taylor & Todd (1995b)
Configuration software	Salespersons	Keil, Beranek & Konsynski (1995)
E-mail	Graduate students	Szajna (1996)
C.A.S.E.	Information technology professionals	Chau (1996)
(Not specified)	Students	Venkatesh and Davis (1996)
Spreadsheet, database, word processor, graphics	Students	Jackson, Chow & Leitch (1997)
Personal computing	PC users	Igbaria et al. (1997)
Debugging tool	Students	Bajaj & Nidimolu (1998)
Configuration software	Salespersons	Gefen & Keil (1998)
Word processing spreadsheet graphics	Users of a Fortune 100 company	Agarwal & Prasad (1999)
Multifunctional workstation	Brokers and sales assistant of financial company	Lucas & Spitler (1999)
Microsoft Windows 3.1	Potential adopters and users in a corporation	Karahanna, Straub & Chervany (1999)
Telemedicine software	Physicians	Hu et al. (1999)
Software maintenance tools	(Not specified)	Dishaw & Strong (1999)
Four different systems in four organizations	Floor supervisors, members of personal financial services	Venkatesh & Davis (2000)
Data and information retrieval	Professionals	Venkatesh & Morris (2000)
Telemedicine software	Physicians	Chau & Hu (2002)

With regard to the measurement of the perceived ease of use dimension, the items most frequently used are as follows:

- Learning to operate (information technology) is easy for me.
- I find it easy to get the (information technology) to do what I want to do.
- The (information technology) is rigid and inflexible to interact with.
- Overall, I find the (information technology) easy to use.

Limitations of the Investigations with a Theoretical Base on the Technology Acceptance Model

A review of the studies that used the Technology Acceptance Model as a predictive model of the use of new technologies allowed us to identify some limitations of the research carried out. The most outstanding are the following:

- Starting with a selection of the sample to validate the model, we consider that the used samples comprise very specific population profiles. In many of the reviewed investigations, a choice of student groups as a target population stands out, a quite common practice for economic reasons. This fact limits the power of generalization of their results to the population as a whole.
- As for the type of technologies that represent the social object of interchange in the examined studies, we notice how all of them center on the prediction of the use of specific technologies. The analysis of office automation computer programs or system development applications must be pointed out. The circumstance described reduces the generality of the model and questions its applicability to any situation regarding the use of technology.
- The investigations carried out propose the existence of two key factors to take into consideration when user attitude toward the use of a certain technology is assessed. However, there is a unanimous

Table 2 Perceived usefulness questions

Perceived Usefulness	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Using (information technology) improves the quality of the work I do	•					•				
Using (information technology) gives me greater control over my work	•									
(Information technology) enables me to accomplish tasks more quickly	•					•	•	•		
(Information technology) supports critical aspects of my job	•									
Using (information technology) increases my productivity	•	•	•	•		•	•	•	•	•
Using (information technology) increases my job performance	•	•	•		•	•	•	•	•	•
Using (information technology) allows me to accomplish more work than would otherwise be possible	•									
Using (information technology) enhances my effectiveness on the job	•	•	•	•			•	•	•	•
Using (information technology) makes it easier to do my job	•			•		•	•	•		
Overall, I find the (information technology) useful in my job	•	•	•		•	•	•	•	•	•

(1) Davis (1989, 1993); (2) Davis et al. (1989); (3) Mathieson (1991); (4) Subramanian (1994); (5) Taylor & Todd (1995a, 1995b); (6) Keil et al. (1995); (7) Szajna (1996); (8) Chau (1996); (9) Jackson et al. (1997); (10) Igbaria et al. (1997)

continued on following page

Table 2. continued

Perceived Usefulness	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
Using (information technology) improves the quality of the work I do		•	•		•					
Using (information technology) gives me greater control over my work			•							
(Information technology) enables me to accomplish tasks more quickly		•	•		•	•	•			•
(Information technology) supports critical aspects of my job										
Using (information technology) increases my productivity	•	•	•	•		•	•	•	•	•
Using (information technology) increases my job performance	•	•					•	•	•	
Using (information technology) allows me to accomplish more work than would otherwise be possible										
Using (information technology) enhances my effectiveness on the job	•		•	•	•	•	•	•	•	•
Using (information technology) makes it easier to do my job			•	•	•	•	•			•
Overall, I find the (information technology) useful in my job	•	•	•	•			•	•	•	
(11) Bajaj & Nidimolu (1998); (12) Gefen & Keil (1998); (13) Agarwal & Prasad (1997, 1999); (14) Lucas & Spitler (1999); (15) Karahanna et al. (1999); (16) Hu et al. (1999); (17) Dishaw & Strong (1999); (18) Venkatesh & Davis (1996, 2000); (19) Venkatesh & Morris (2000); (20) Chau & Hu (2002)										

agreement among authors when they concur that it is necessary to include other components in the model so they can contribute to the explanation of the variance in the use of technology. This suggests that it would be convenient to thoroughly study the dimensionality within the technology attitude concept, redefining and improving some of the variables the various dimensions are made of, and also attempting to identify new ones.

Technology Readiness Model

The limitations encountered by the Technology Acceptance Model suggest a desire for an exhaustive and systematic research on the multidimensionality of the technology attitude concept, bearing in mind perceptions different types of users may have about the use of a wide range of products with a relevant technological content. That is the purpose of investigations carried out by Parasuraman and Colby (1997, 1998, 2001), which ended with the creation of the Technology Readiness Index (TRI) as a tool of measurement of the degree of technology readiness or attitude.

This section deals with the detailed description of the methodology followed by these authors in order to build up a measurement scale of technology attitude, which we consider comprises and enlarges previous formulations. A comparison of the structural model of attitude, seen both from the inside of the Technology Acceptance Model and also the Technology Readiness Model, allowed us to come to such a conclusion.

Description of the Technology Readiness Model

The concept of technology readiness was first introduced in marketing literature by Parasuraman and Colby (1997) and refers to a combination of beliefs related to technology that collectively determine the inclination in a customer, employee, or executive to adopt new technologies in order to reach his or her objectives, both at work and during leisure time. The authors use the term “technology readiness” as a synonym of

Table 3. Perceived ease of use questions

Perceived Ease of Use	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
I find (information technology) cumbersome to use	•					•		•		
Learning to operate (information technology) is easy for me	•	•	•	•	•	•	•	•	•	
Interacting with the (information technology) is often frustrating	•					•				
I find it easy to get the (information technology) to do what I want to do	•	•	•			•	•		•	•
The (information technology) is rigid and inflexible to interact with	•			•			•	•		•
It is easy for me to remember how to perform tasks using the (information technology)	•									
Interacting with the (information technology) requires a lot of mental effort	•					•				•
My interaction with the (information technology) is clear and understandable	•					•				
I find it takes a lot of effort to become skillful at using the (information technology)	•	•	•	•			•	•		
Overall, I find the (information technology) easy to use	•	•	•	•	•	•	•		•	•

(1) Davis (1989, 1993); (2) Davis, Bagozzi and Warshaw (1989); (3) Mathieson (1991); (4) Subramanian (1994); (5) Taylor and Todd (1995a, 1995b); (6) Keil, Beranek and Konsynski (1995); (7) Szajna (1996); (8) Chau (1996); (9) Jackson, Chow and Leitch (1997); (10) Igbaria, Zinatelli et al (1997)

Perceived Ease of Use	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
I find (information technology) cumbersome to use										
Learning to operate (information technology) is easy for me		•			•	•				•
Interacting with the (information technology) is often frustrating										
I find it easy to get the (information technology) to do what I want to do	•		•	•		•	•	•	•	•
The (information technology) is rigid and inflexible to interact with			•			•	•	•		•
It is easy for me to remember how to perform tasks using the (information technology)			•							
Interacting with the (information technology) requires a lot of mental effort	•					•	•		•	•
My interaction with the (information technology) is clear and understandable		•						•	•	
I find it takes a lot of effort to become skillful at using the (information technology)						•				•
Overall, I find the (information technology) easy to use	•	•	•	•	•	•	•	•	•	•

(11) Bajaj & Nidimolu (1998); (12) Gefen & Keil (1998); (13) Agarwal & Prasad (1997, 1999); (14) Lucas & Spitler (1999); (15) Karahanna et al. (1999); (16) Hu et al. (1999); (17) Dishaw & Strong (1999); (18) Venkatesh & Davis (1996, 2000); (19) Venkatesh & Morris (2000); (20) Chau & Hu (2002)

attitude toward technology, the latter being the most frequently used term in the bibliography relevant to this subject matter.

From the concept of technology readiness as adopted, and following Churchill's (1979) directives, in order to create measurement scales, the authors identified those dimensions belonging to the domain of the construct. Parasuraman and Colby (1997, 1998, 2001) formulated a model in which technology readiness is enhanced by two variables and inhibited by the other two, as can be seen in Figure 4.

The model of the technology attitude structure as presented by Parasuraman and Colby (1997, 1998, 2001) is consistent with other authors' proposals that, as we have seen in previous sections, defended the simultaneous coexistence of favorable and unfavorable beliefs toward technology inside the user's mind.

The results of the investigation carried out by Mick and Fournier (1998) back Parasuraman and Colby's (1997, 1998, 2001) conclusions as well, as far as the dimensionality of the proposed concept is concerned. The authors, on the grounds of an extensive qualitative investigation on the reactions of people toward technology, identified eight technology paradoxes the user must face: control/chaos, freedom/enslavement, new/obsolete, competence/incompetence, efficiency/inefficiency, fulfills/creates needs, assimilation/isolation, and engaging/disengaging.

Apart from the bibliographic review, Parasuraman and Colby (1997, 1998, 2001) developed an extensive four-year research program about attitudes toward technology that included another type of qualitative analysis. The authors examined in detail the results of focus group interviews and case studies carried out with clients in various sectors (financial services, telecommunications, e-commerce, education, tourism, etc.) about beliefs related to the use of technology. The results of such studies are consistent with the aforementioned idea with regard to the simultaneous existence of enhancing and inhibitory factors of technology readiness. Flexibility, convenience, efficiency, and enjoyment were some of the emphasized advantages of technology, as well as disadvantages, security problems, obsolescence risks, impersonal treatment, and difficulty of use, to name but a few.[1]

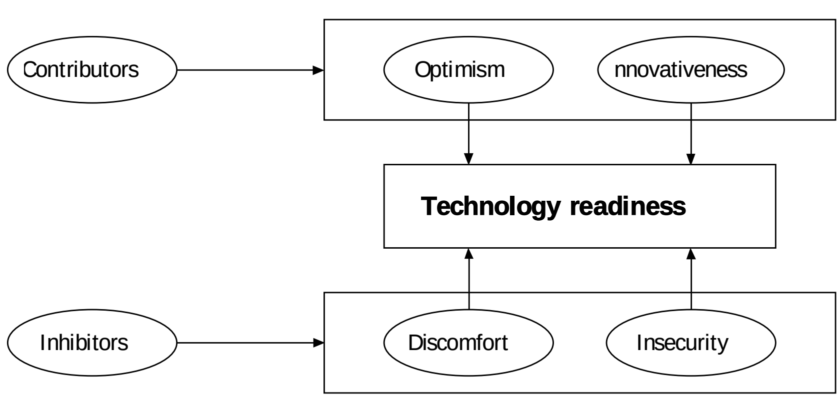
The dimensions of the Technology Readiness Model are described with maximum detail in the following headings.

Optimism

Optimism dimension was defined by Parasuraman and Colby (2001) as "a positive view of technology and a belief that it offers people increased control, flexibility and efficiency in their lives" (p. 32). In particular, the 10 indicators that Parasuraman suggests to measure this latent variable reveal a series of beliefs related to the benefits of technology:

- Acquisition of the latest technology is very convenient
- Computers extend the opening hours of commercial outlets
- Computers are easier to deal with than other people
- Technology increases working efficiency

Figure 4. Drivers of technology readiness (Source: Parasuraman, 2000a)



The positive perception of technology changes with age. In particular, people over 65 years old are more sceptical about the advantages of the use of the new technologies, as they consider them to be overestimated and often produce disappointing results.

Nevertheless, no significant differences appear as far as optimism toward technology among individuals of different sex is concerned. Both men and women show a rather positive vision toward technology, albeit with certain reservations. That is what can be drawn out from the remarks both genders make about the difficulty of doing without human contact when a service with a high technology component is supplied.

Innovativeness

The innovativeness dimension refers to "a tendency to be a technology pioneer and thought leader" (Parasuraman & Colby, 2001, p. 38). People who score high in this dimension are the first ones to adopt technologies and tend to learn to use them by themselves.

Seven indicators are suggested to measure this latent variable (Parasuraman, 2000b), some of them strongly connected with age and sex.

So the older people are, the lower the innovativeness. In this respect, it seems that the tendency to be an opinion leader stands out in individuals in approximately the 30- to 40-year bracket. People who are 50 years and over tend to lag behind when acquiring and using the latest technologies in which they are interested. Also, from that age on, the challenge to fully understand high-tech mechanisms is much less enjoyable, and right then, the aid and assistance of others become necessary. For people living beyond 65 years of age, as they grow older, they are less receptive to learn anything about new technology trends.

In general, innovativeness is much more developed in men than in women. There is a higher possibility of men not lagging behind in the implementation of the most advanced technologies that provide them with some kind of advantage and the desire to understand how they operate. On the whole, it is men who are among the first within their circle of friends to acquire the latest technological breakthroughs. They can usually grasp the new technology services and products without anyone else's assistance. Actually, they are the ones who give advice to others in this matter.

Discomfort

The discomfort dimension can be defined as "a perceived lack of control over technology and a feeling of being overwhelmed by it" (Parasuraman & Colby, 2001, p. 41). A definition like this could lead us to confuse it with the perceived behavioral control factor appearing in the Planned Behavior Model. Nonetheless, once we have analyzed the 10 items related to this dimension (Parasuraman, 2000b), we consider that what it really represents is a lack of a user's comfort when handling technology, which derives from the perceived difficulty of use and the wariness its use arouses.

Individuals with high marks in this dimension reckon technology is something too tricky and difficult to be really useful. They feel overwhelmed because they think technology is not for ordinary people since it demands a continuous updating effort to acquire the necessary knowledge enabling them to take advantage of the latest innovations. This view about technology is more widespread among women as well as elderly people. Thus, we can observe that people aged 65 years and older consider the effort to manage the new technologies not worth doing in comparison with the profit they get in return. Such effort is far more justified by people under 30.

People who score high in this dimension feel a lack of effectiveness in technical assistance services. In particular, they consider the technical assistance they are provided with, either by people or handbooks, not useful inasmuch as things are not explained in intelligible terms. Actually, when interacting with technical assistance staff, they feel as if they were in a disadvantageous position compared with somebody who knows more. Like in the previous case, this set of beliefs is mostly shared by females, even more so as they grow older.

Insecurity

Finally, the insecurity dimension can be defined as "distrust of technology and skepticism about its ability to work properly" (Parasuraman & Colby, 2001, p. 44). There is a certain relationship between the discomfort dimension and some of the nine items used to measure insecurity dimension. Their main difference is that indicators of the insecurity dimension allow us to evaluate specific aspects of technology-based transactions instead of a discomfort with technology, generally speaking.

People who dislike e-business are expected to get high marks in this dimension. They fear that their personal data might be stolen if stored inside a computer. In contrast to the other dimensions of the technology readiness concept mentioned previously, and with regard to beliefs hereby assessed, no differences in age or sex can be observed.

Scales of Technology Readiness

Starting by defining the domain of the construct derived from their exploratory works, Parasuraman and Colby (1997) set an initial batch of 44 items to measure the degree of technology readiness. Next, and having carried out several empirical studies (details are shown in Table 2.4), the scale was filtered until they obtained a measurement tool comprised of 36 items (see Table 2.5) where nearly 50% of them corresponded to positive beliefs toward technology and the rest with negative beliefs.

The decision to measure technology readiness or attitude with beliefs related to use of technology in consideration was already supported by various authors (Cowles, 1989, 1991; Cowles & Crosby, 1990; Eastlick, 1996), who defended the existence of a strong and significant correlation between beliefs and attitude toward technology as shown in Figure 5.

The measurement scale eventually put forward by Parasuraman (2000b) overcomes earlier versions' deficiencies, thus achieving its applicability to different contexts of use of technology. In this particular case, previous investigations had focused on the analysis of the degree of readiness toward the use of computers or the Internet, not toward new technologies as a whole. Also, the scale had been validated by means of very specific samples, which restricted the power to generalize results to all the population.

Parasuraman (2000b) presents his scale of technology readiness as an suitable alternative to assess clients, employees, or executives' attitudes toward the adoption of the new technologies as a whole, with a view to succeeding in their jobs and leisure time. The procedure followed to construct this scale guarantees its validity. Thus, as far as its content validity is concerned, we have to point out that the generated sample was adequate and representative enough of the contents and dimensions of the studied concept by means of a review of written accounts of the focus group interviews previously held (Parasuraman & Colby, 1997) and also the carrying out of new brainstorming sessions. In this way, the initial 28-item scale proposed by Parasuraman and Colby (1997), was enlarged with 38 additional items. Later on, an iterative scale-purification process following Churchill's (1979) methodology was achieved, with a final result of a 36-item measurement scale.

As for the operational features of the scale, for every proposition, interviewees had to express their degree of assent within Likert-type scale of five points (from "strongly disagree" to "strongly agree").

Comparison Between the Technology Readiness Model and the Technology Acceptance Model

We finish the chapter with a comparison between the two predictive models of use of technology previously described. In doing this, we intend to explain why we have chosen the Technology Readiness Model as a theoretical basis to develop a scale of technology attitude.

Starting with the definition of the technology attitude concept that these models respectively formulate, we can observe that the Technology Readiness Model coins a generic name for the type of technologies that constitute the social object of research interchange, not being centered around the analysis of attitude toward a particular technology, as occurs in the case of the Technology Acceptance Model.

With regard to the specification of the concept's dimensions by both models, we notice certain coincidences in some dimensions' contents. Thus, the perceived usefulness dimension (the Technology Acceptance Model) is contained within a definition of the optimism dimension (the Technology Readiness Model), the latter gaining a much wider perspective of the benefits obtained through the use of technologies. The perceived ease of use dimension (the Model of Technology Acceptance) comprises in its definition the contents of the discomfort dimension (the Technology Readiness Model), although the latter includes an evaluation of other overall aspects of technology-based transactions that inhibit its use. The Technology Readiness Model also covers two other attitude dimensions—innovativeness and insecurity—incorporating them into a much wider perspective of all the factors that either encourage or inhibit the use of technology.

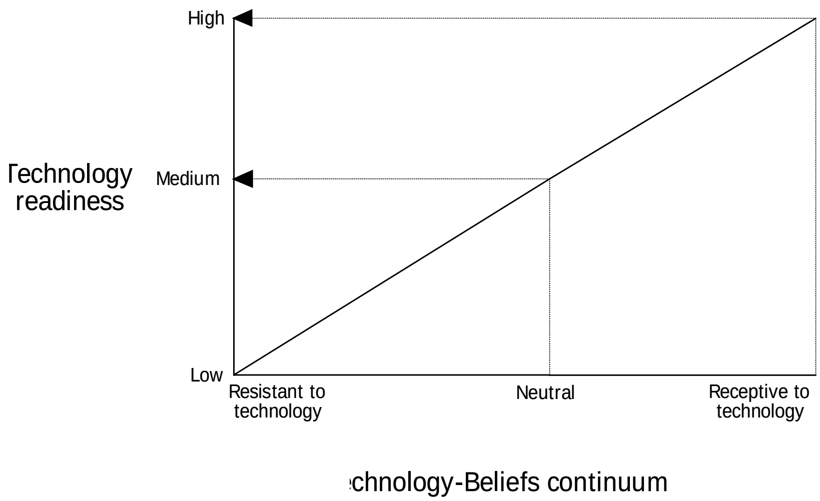
Table 4. Comparison of Parasuraman and Colby (1997) study with other TRI replication studies

Study	**Parasuraman and Colby (1997)**	**Parasuraman and Colby (1998) (Study 1)**	**Parasuraman and Colby (1998) (Study 2)**	**Parasuraman (2000b)**
Data collection study sample	Clients of a firm offering a variety of services related to financial assistance for college students and young college graduates (up to age 35)	Clients of a mortgage bank	Clients of an online service provider	Adults (18 years or older)
Sample size	1,200 individuals			1,000 individuals
Response scale	Five-point scale	Five-point scale	Five-point scale	Five-point scale
Questionnaire administration	Mail survey and Web survey			Computer-assisted telephone survey
Analysis techniques	Test of reliability (internal consistency) Confirmatory factor analysis	Test of reliability (internal consistency) Confirmatory factor analysis	Test of reliability (internal consistency) Confirmatory factor analysis	Test of reliability (internal consistency) Confirmatory factor analysis
Basis for initial number of factors extracted	Factors with eigen values greater than 1	Parasuraman and Colby's (1997) four-dimensional structure	Parasuraman and Colby's (1997) four-dimensional structure	Parasuraman and Colby's (1997) four-dimensional structure
Reliability	Cronbach alpha (0.72-0.82) Item-Total Correlation Principal-axis factor analysis followed by varimax rotation	Cronbach alpha (weaker than in Parasuraman and Colby (1997) study for the four subscales) Item-Total Correlation Principal-axis factor analysis followed by varimax rotation	Cronbach alpha (weaker than in Parasuraman and Colby (1997) study for the four subscales) Item-Total Correlation Principal-axis factor analysis followed by varimax rotation	Cronbach alpha (0.74-0.81) Item-Total Correlation Principal-axis factor analysis followed by varimax rotation
Validity	Content validity Construct validity	Content validity Construct validity	Content validity Construct validity	Content validity Construct validity
Final number of dimensions	4	4	4	4
Original TRI item retained	28 items	26 items	16 items	21 items
Number of items obtained	28 items: Optimism (10), Innovativeness (5), Discomfort (8), and Insecurity (5)	26 items	16 items	36 items: Optimism (10), Innovativeness (7), Discomfort (10), and Insecurity (9)

Table 5. Items used to measure technology readiness (1) Parasuraman and Colby (1997); Parasuraman (2000b); (2) Parasuraman (2000b)

TR STATEMENTS	ITEMS	STUDY
OPTIMISM STATEMENTS	Technology gives people more control over their daily lives	(1)
	Products and services that use the newest technologies are much more convenient to use	(1)
	You like the idea of doing business via computers because you are not limited to regular business hours	(1)
	You prefer to use the most advanced technology available	(1)
	You like computer programs that allow you to tailor things to fit your own needs	(1)
	Technology makes you more efficient in your occupation	(1)
	You find new technologies to be mentally stimulating	(1)
	Technology gives you more freedom of mobility	(2)
	Learning about technology can be as rewarding as the technology itself	(2)
	You feel confident that machines will follow through with what you instructed them to do	(2)
INNOVATIVE STATE-MENTS	Other people come to you for advice on new technologies	(1)
	It seems your friends are learning more about the newest technologies than you are [reverse scored]	(1)
	In general, you are among the first in your circle of friends to acquire new technology when it appears	(1)
	You can usually figure out new high-tech products and services without help from others	(1)
	You keep up with the latest technological developments in your areas of interest	(2)
	You enjoy the challenge of figuring out high-tech gadgets	(2)
	You find you have fewer problems than other people in making technology work for you	(2)
INSECURE STATEMENTS	The human touch is very important when doing business with a company	(2)
	When you call a business, you prefer to talk to a person rather than a machine	(2)
	If you provide information to a machine or over the Internet, you can never be sure it really gets to the right place	(1)
	You do not consider it safe giving out a credit card number over a computer	(1)
	You do not consider it safe to do any kind of financial business online	(1)
	You worry that information you send over the Internet will be seen by other people	(1)
	You do not feel confident doing business with a place that can only be reached online	(1)
	Any business transaction you do electronically should be confirmed later with something in writing	(1)
	Whenever something gets automated, you need to check carefully that the machine or computer is not making mistakes	(2)
DISCOMFORT STATEMENTS	Technical support lines are not helpful because they don't explain things in terms you understand	(1)
	Sometimes you think that technology systems are not designed for use by ordinary people	(1)
	There is no such thing as a manual for a high-tech product or service that's written in plain language	(1)
	When you get technical support from a provider of a high-tech product or service, you sometimes feel as if you are being taken advantage of by someone who knows more than you do	(1)
	If you buy a high-tech product or service, you prefer to have the basic model over one with a lot of extra features	(2)
	It is embarrassing when you have trouble with a high-tech gadget while people are watching	(2)
	There should be caution in replacing important people-tasks with technology because new technology can break-down or get disconnected	(2)
	Many new technologies have health or safety risks that are not discovered until after people have used them	(2)
	New technology makes it too easy for governments and companies to spy on people	(2)
	Technology always seems to fail at the worst possible time	(2)

Figure 5. Links between technology beliefs and technology readiness (Source: Parasuraman and Colby, 2001)



With relation to all the items used to measure the technology attitude by means of a theoretical base on the analyzed models, we can remark that contents that represent items most frequently used to measure latent variables perceived usefulness and perceived ease of use are present in the scale of technology readiness.

All in all, we can consider that validations carried out of the Technology Readiness Model possess a higher applicability to different contexts of use of technology than those of the Technology Acceptance Model. That is why we estimate appropriately the use of the Technology Readiness Model as a theoretical basis to develop a measurement scale of technology attitude.

REFERENCES

Agarwal, R., & Prasad, J. (1999). Are individual differences germane to the acceptance of new information technologies., *Decision Sciences*, *30*(2), 361–391.

Ajzen, I. (1985). From intentions to actions: A theory of planned behavior. In J. Kuhl, & J. Beckmann (Eds.), *Action control: From cognition to behavior* (pp. 11–39). New York: Springer Verlag.

Ajzen, I. (1991). The theory of planned behavior., *Organizational Behavior and Human Decision Processes*, *50*, 179–221.

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Englewood Cliffs, NJ: Prentice Hall.

Ajzen, I., & Madden, T.J. (1986). Prediction of goal directed behaviours: Attitudes, intentions, and perceived behavioral control., *Journal of Experimental Social Psychology*, *22*, 453–474.

Bajaj, A., & Nidimolu, S.R. (1998). A feedback model to understand information system usage. *Information and Management*, *33*, 213–224.

Chau, P.Y. (1996). An empirical investigation on factors affecting the acceptance of CASE by systems developers., *Information and Management*, *30*, 269–280.

Chau, P.Y., & Hu, P.J. (2002). Examining a model of information technology acceptance by individual professionals: An exploratory study., *Journal of Management Information Systems*, *18*(4), 191–229.

Churchill, G.A. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research*, *16*, 64–73.

Cowles, D. (1989). Consumer perceptions of interactive media. *Journal of Broadcasting and Electronic Media*, *33*(Winter), 83–89.

Cowles, D. (1991). Retail bankers: Marketing may not be enough. *Bank Marketing*, *23*, 22–25.

Cowles, D., & Crosby, L.A. (1990). Consumer acceptance of interactive media in service marketing encounters. *Service Industries Journal*, *10*, 521–540.

Davis, F.D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319–340.

Davis, F.D. (1993). User acceptance of information technology: System characteristics, user perceptions, and behavioural impacts. *International Journal of Man Machine Studies*, *38*, 475–487.

Davis, F.D., Bagozzi, R.P., & Warshaw, P.R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, *35*(8), 982–1003.

Dishaw, M.T., & Strong, D.M. (1999). Extending the technology acceptance model with task-technology fit constructs. *Information and Management*, *36*, 9–21.

Eastlick, M.A. (1996). Consumer intention to adopt interactive teleshopping. In *Marketing science institute working paper (August)* (pp. 96–113). Cambridge: Marketing Science Institute.

Gattiker, U.E. (1984). Managing computer-based office information technology: A process model for management. In H. Hendrick, & O. Brown (Eds.), *Human factors in organizational design* (pp. 111–132). Amsterdam: North-Holland.

Gefen, D., & Keil, M. (1998). The impact of developer responsiveness on perceptions of usefulness and ease of use: An extension of technology acceptance model. *The DATA BASE for Advances and Information Systems*, *29*(2), 35–49.

Hu, P.J., et al. (1999). Examining the technology acceptance model using physician acceptance of telemedicine technology. *Journal of Management Information Systems*, *16*(2), 91–112.

Igbaria, M., et al. (1997). Personal computing acceptance factors in small firms: A structural equation model. *MIS Quarterly*, September, 279–302.

Jackson, C.M., Chow, S., & Leitch, R.A. (1997). Toward an understanding of behavioural intention to use an information system. *Decision Sciences*, *28*(2), 357–389.

Karahanna, E., Straub, D.W., & Chervany, N.L. (1999). Information technology adoption across time: A cross-sectional comparison of pre-adoption and post-adoption beliefs. *MIS Quarterly*, *23*(2), 183–213.

Keil, M., Beranek, P.M., & Konsynski, B.R. (1995). Usefulness and ease of use: Field study evidence regarding task considerations. *Decision Support Systems*, *13*, 17–91.

Lucas, H.C., & Spitler, V.K. (1999). Technology use and performance: A field study of broker workstations. *Decision Sciences*, *30*(2), 291–311.

Mathieson, K. (1991). Predicting user intentions: Comparing the technology acceptance model with the theory of planned behavior. *Information Systems Research*, *2*(3), 173–191.

Mick, D.G., & Fournier, S. (1998). Paradoxes of technology: Consumer cognizance, emotions and coping strategies. *Journal of Consumer Research*, *25*, 123–147.

Morales, J.F., Rebolloso, E., & Moya, M. (1996). Modelos combinatorios del cambio de actitudes. In J.F. Morales, et al (Eds.), *Psicología social* (pp. 555–566). Madrid: McGraw-Hill.

Parasuraman, A. (2000a). Technology's role in service delivery: Implications for measuring and improving service quality. In E. Ortega, L. González, & E. Pérez (Eds.), *III forum internacional sobre las ciencias, las técnicas y el arte aplicadas al marketing. Academia y profesión*. Madrid: Universidad Complutense de Madrid.

Parasuraman, A. (2000b). Technology readiness index (TRI). A multiple-item scale to measure readiness to embrace new technologies. *Journal of Service Research*, *2*(4), 307–320.

Parasuraman, A., & Colby, C. (1997). Correlates and consequences of consumer attitudes toward new technologies: Implications for marketing technology-based services. Retrieved from www.rockresearch.com/Technology_Readiness/tq01b/html

Parasuraman, A., & Colby, C. (1998). A scale for measuring customers' technology readiness: Replication, refinement, and implications for service organizations. Proceedings of the 1998 Frontiers in Services Conference, Nashville, Tennessee. Cited in A. Parasuraman (2000b), Technology readiness index (TRI). A multiple-item scale to measure readiness to embrace new technologies. *Journal of Service Research*, *2*(4), 307–320.

Parasuraman, A., & Colby, C.L. (2001). *Techno-ready marketing: How and why your customers adopt technology*. New York: The Free Press.

Subramanian, G.H. (1994). A replication of perceived usefulness and perceived ease of use measurement. *Decision Sciences*, *25*(5/6), 863–874.

Szajna, B. (1996). Empirical evaluation of the revised technology acceptance model. *Management Science*, *42*(1), 85–92.

Taylor, S., & Todd, P.A. (1995a). Understanding information technology usage: A test of competing models. *Information Systems Research*, *6*(2), 144–176.

Taylor, S., & Todd, P.A. (1995b). Assessing IT usage: The role of prior experience. *MIS Quarterly*, December, 561–570.

Venkatesh, V., & Davis, F.D. (1996). A model of antecedents of perceived ease of use: Development and test. *Decision Sciences*, *27*(3), 451–481.

Venkatesh, V., & Davis, F.D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science*, *46*(2), 186–204.

Venkatesh, V., & Morris, M.G. (2000). Why do not men ever stop to ask for directions? Gender, social influence, and their role in technology acceptance and usage behaviour. *MIS Quarterly*, *24*(1), 115–139.

KEY TERMS

Cognitive Dissonance Theory: A theory that discusses how the tension that may result from having two conflicting thoughts at the same time, or from engaging in behavior that conflicts with one's beliefs, or from experiencing apparently conflicting phenomena, can impact actions and/or decision-making.

Expectancy-Value Theory: Explains and predicts individual's attitudes toward objects and actions but is also often the basis for other theories to explain perception and behaviors.

Reasoned Action Model: A model to predict and understand human behaviour determinants.

Self-Perception Theory: Discusses that an account of attitude change can occur by observing our own behavior and concluding what attitudes must have caused them.

Technology Acceptance Model: A model that discusses that perceived usefulness and ease of use are two important factors in influencing the acceptance of a technology by a user.

ENDNOTE

1 The authors also noted that some beliefs toward technology overlapped with criteria that clients use when assessing the quality of service, so the attributes of the SERVQUAL scale (Parasuraman, Zeithaml & Berry, 1988), a tool to measure perceived quality widely cited in marketing literature for its application in a wide range of tertiary activities with very different features.

The Detection of Abnormal Breathing Activity by Vision Analysis in Application to Diagnosis of Obstructive Sleep Apnea

Ching-Wei Wang
University of Lincoln, UK

Andrew Hunter
University of Lincoln, UK

INTRODUCTION

Obstructive Sleep Apnea is increasingly seen as a common and important condition, contributing to sleep disturbance and consequential daytime sleepiness. This has potentially serious consequences for the individual, employers, and society as a whole. Apnea refers to short spells when breathing stops. In obstructive sleep apnea, the throat constricts during sleep, preventing breathing; the apnea episode often ends with a loud snore and/or gasp. Such an event is sufficient to open the throat muscles to allow breathing, and the patient usually falls asleep again so quickly that the event is not remembered. This cycle repeats itself throughout the night as the muscles relax and the throat blocks again, and the frequency of the episode is used to determine the severity of the syndrome.

A wide range of parameters, including EEG (electroencephalogram) sleep staging, snoring, changes in airway resistance, airflow, and inspiratory effort, as well as oxygen saturation and body movement during "normal sleep" in a Sleep Lab, lead to an understanding of the pathophysiology of sleep apnea (Flemons et al, 2003). According to recent research findings (Bennett, Langford, Stradling & Davies, 1998; Kingshott et al., 2000; Pepperell, Davies & Stradling, 2002), the best predictors of morbidity in individual patients, as assessed by improvements with continuous positive airway pressure therapy (CPAP), are nocturnal oxygen saturation and movement during sleep. Moreover, Siven, Kornecki, and Schonfeld (1996) indicate that the results from traditional polysomnography are highly correlated with the video test results.

Video monitoring has been adopted to assist diagnosis on obstructive sleep apnea. However, video monitoring and interpretation are less well developed, both from a recording and analysis viewpoint, due to the relative computational complexity of video analysis and the significant technical challenges. These include the lighting restriction involved with natural night vision and low illumination, including a lack of no color information; the obscuration of the body by a cover; the variation of human size, shape, and behavior; and the massive volume of video and audio data. The diagnosis of sleep disturbances requires a review of substantial amounts of data (including video footage) by clinicians. There is therefore a need for automated monitoring of human breathing activity to diagnose sleep disorders.

BACKGROUND

Current breathing monitoring techniques can be categorized into two types: invasive and noninvasive. The invasive approaches include Polysomnography; the Belt (Hyun Medics, 2006) or Strap (Svetlana et al., 2005), which track changes in body circumference during the respiratory cycle; Stick-on Electrodes as a heart-respiratory monitor (David et al., 2005); and Nasal Temperature Probes (Storck, Karlsson, Ask & Loyd, 1996). Published noninvasive techniques include Audio Analysis to monitor tidal volumes from human breathing activity (Amin, Cigada, Fordyce & Camporesi, 1993), Vibration Sensors (David et al., 2003; Randall, 1995), and Thermal Imaging (Murthy, Pavlids & Tsiamyrtzis, 2004; Zhu, Fei & Pavlidis, 2005).

However, these techniques for monitoring breathing activities have various limitations in application to diagnosis of sleep disorders. The obtrusive nature of invasive monitoring equipment can disturb sleep and therefore compromise results. The thermistors used in Polysomnography sense differences in temperature; however, they do not have a linear relationship with true airflow. Consequently, the thermistors may not be sensitive for detecting hypopneas (Flemons et al., 2003). Nasal pressure has a linear approximation of airflow but can produce false-positive events and low quality signals if patients are mouth breathing (Flemons et al., 2003). Regarding strap systems, if the tension on the strap is not calibrated, the system will not track the respiration motion correctly, so adjustment may be necessary. Moreover, measurements on patients with shallow and abdominal breathing patterns may fail because the sensor cannot track adequately in a reproducible manner if the chest displacements during normal breathing and breath-hold are not distinctly different. Concerning thermal imaging techniques (Murthy et al., 2004; Zhu et al., 2005), there are position limitations and geometric constraints for targeting faces, which will not be applicable for unpredicted human sleep behavior. As a result, an investigation of methods for monitoring human breathing activity is crucial for diagnosis of sleep disorders.

METHODS

In this chapter, we develop a nonintrusive monitoring technique without geometric constraints using infrared video information for identifying abnormal breathing activity in application to diagnosis of obstructive sleep apnea. The proposed technique allows patients to sleep on their back or side, with or without facing the camera. The experiments show that the proposed method obtains promising results in recognizing abnormal breathing events and identifying general body movements. Moreover, the experiments also show that the proposed approach is able to recognize events successfully for patients with shallow and abdominal breathing patterns.

Objectives

The aim of our research is to support the diagnosis of obstructive sleep apnea. Our primary objective is to detect abnormal breathing episodes based on vision analysis. This also requires us to distinguish breathing movements from other body movements.

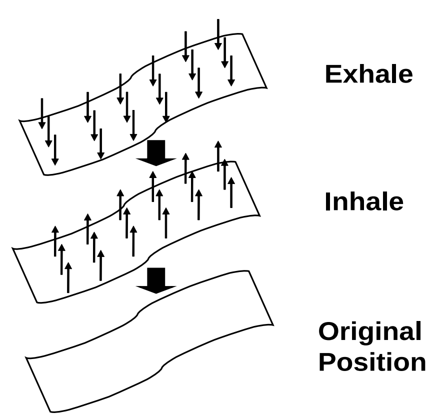
Technical Analysis in Breathing Behavior

In order to recognize abnormal breathing activities, it is necessary to differentiate between general movements and breathing, allowing further identification of normal breathing status and abnormal breathing status. Thus, we first analyze human breathing behavior in comparison with general body movement and find two important features of the breathing activity. The first is cyclic motion; the elements of the entire surface move forward and backward approximately to their previous position in a breathing cycle (see Figure 1). In contrast, elements tend to move toward different positions for general body movements. Second, considering the rate of movement, breathing is a relatively *slow-motion* activity, compared with other body movements. Hence, in order to detect breathing activity, it is necessary to observe differences across several seconds of video.

Design

Following the previous analysis, we attempt to distinguish breathing from general body movements based on variations in movements. As the breathing activity is relatively slow, we use an adaptive background model to identify motion; this allows relatively subtle motion to be detected and accumulated within a period of time.

Figure 1. Cyclic moving flow in a breath



In addition, the duration and degree of motion are used as criteria to assist in distinguishing breathing motion from other motion. Furthermore, a continuously updated 2D breathing template is created for identifying abnormal breathing events.

Adaptive Background Model

The adaptive background model is crucial for distinguishing general body movement and breathing movement. The model updates slowly, and so can determine if the surface moves back to the previous position. Breathing movements produce detectable cycles in the number of pixels different from the background image. The adaptive background model *B(x, y)* is formulated as follows:

$$B_i(x,y) = B_{i-1}(x,y) + \begin{cases} \eta & if \quad P_i(x,y) - B_{i-1}(x,y) > 1 \\ -\eta & if \quad P_i(x,y) - B_{i-1}(x,y) < -1 \\ 0 & otherwise \end{cases} \tag{2}$$

where $P_i(x, y)$ is the intensity value at time *i*; $B_{i-1}(x, y)$ is the adaptive background at time *i-1*; and $B_i(x, y)$ is the updated adaptive background at time *i*.

Continuously Updated 2D Breathing Template

While the body is in general breathing mode, a 2D breathing template is built and continuously updated over time. The template represents general breathing movement. This novel technique is proposed to both assist classification of the current movement as a breath activity or general body movement, and further to determine if the breathing activity is an abnormal one. Figure 2 illustrates how the 2D breathing template is utilized by matching shapes of current movement and the 2D breathing template to determine the class of the behavior (i.e., breathing movement or general body movement).

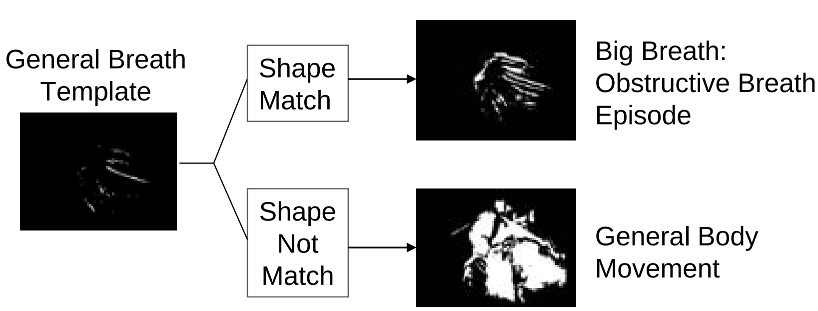
Template Construction. The template represents a combination of shapes of normal breathing movements. It is continuously built over time, and the development does not stop until the status changes to "not general breathing." Figure 3 illustrates the building procedure.

Template Match. As patients with shallow breathing patterns tend to have low-quality breathing templates, a template quality index is created to decide whether the template is mature enough to use. In addition, the degree of movement (*M*) and the duration of the movement (*D*) are employed to eliminate the occurrences of general body movement. The proposed matching method compares the breath template and the shape of the current movement. Given the breathing template $T_i(x, y)$ and the shape of current movement $S_i(x, y)$, the difference between current frame and the adaptive background image, at time *i*, an index *E* is the ratio W_2/W_1, where W_1 is the number of pixels of overlap between the template and the current shape of movement, and where W_2 is the number of pixels that are in the current shape not in the template (see Equations 3, 4, and 5). The index *E* is then used to determine the type of current event (i.e., Apnea or general body movement) (see Table 1).

$$E = W_2 / W_1 \tag{3}$$

$$W_1 = \left| T_i(x,y) \cap S_i(x,y) \right| \tag{4}$$

Figure 2. Breathing template for classification on current activity



$$W_2 = \left| \left(-T_i(x,y) \right) \cap S_i(x,y) \right| \qquad (5)$$

Experimental Setup

Two SONY infrared camcorders (DCR-HC-30E) are utilized to capture a number of video clips from three shooting angles. The infrared video frames were acquired at 15 frames per second, with resolution of 320*240. The video and audio data are first captured with the WMP9 compression algorithm to minimize storage size, and then decompressed for off-line analysis. Moreover, in order to simulate the environment for diagnosis on sleeping disorders, there was no visible lighting in the filming room, and the subjects were partially covered by a sheet. Furthermore, the experimental data were collected from two subjects with mainly three different postures (i.e., lying on the back, sleeping on one side and facing the camera, sleeping on the other side with back facing the camera). The data were also collected on different days with the subjects wearing different clothing. Activities such as general breathing, obstructive apnea, and body movement were simulated by the subjects.

Experimental Results

To evaluate the proposed method for monitoring human breathing activity, the experimental video contents are first analyzed by human observation, defining the real events and marking the frame number of the beginning and ending point for an individual event to produce a reference standard. To facilitate that, the observer was allowed to play the video in slow motion. The results of these analyses are lists of events with the associated frame numbers.

The analysis outputs of the proposed method on sequences of event type and the frame numbers of cutting points among events are then compared to the reference standard. Figure 4 presents the experimental results of the proposed method and manual observation on fifteen video clips, which demonstrates that the proposed vision analysis model achieves high accuracy in recognizing abnormal breathing activities and general body movements.

Figure 5 illustrates the analysis results of three video samples. In the upper area, the motion degrees in temporal space are presented with the number of episodes marked. The lower area shows system procedures. For example, there are seven episodes in the first sample, representing Episode 1: continuously construct a breathing template; Episodes 2 and 3: use the template built in episode 1 to do template matching and define the event as obstructive apnea; Episode 4: build a new breathing template; Episode 5: use the template built in episode 4 for template matching and define the event as a general body movement event; Episode 6: build a new breathing template; Episode 7: use the template from episode 6 and define the event as obstructive apnea.

Figure 3. Template construction

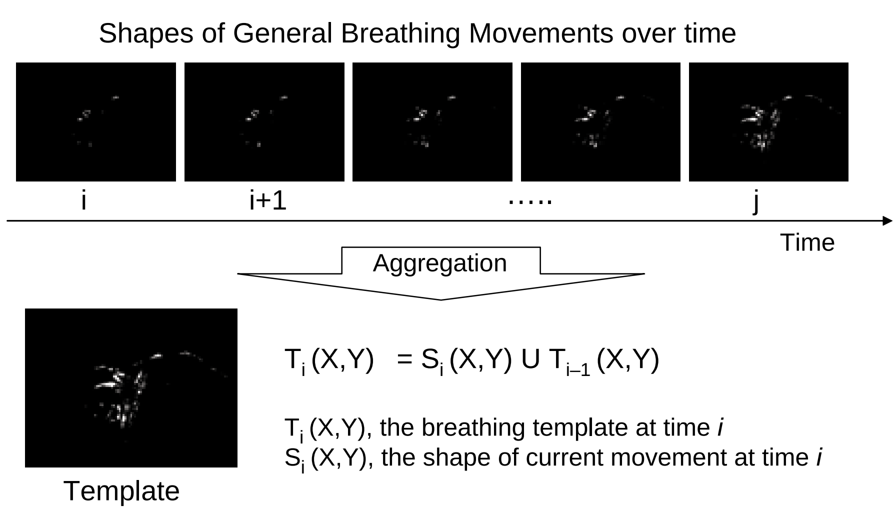


Table 1. Algorithm: System flow

			
Event: Apnea *E* = 8.30	Event: Body *E* = 47.47	Event: Body *E* = 17.20	Event: Apnea *E* = 8.86
			■ : W_1 ■ : W_2 ■ : $T_i(x,y) \wedge \sim S_i(x,y)$
Event: Breath *E* = 0.95	Event: Body *E* = 21.11	Event: Apnea *E* = 3.06	

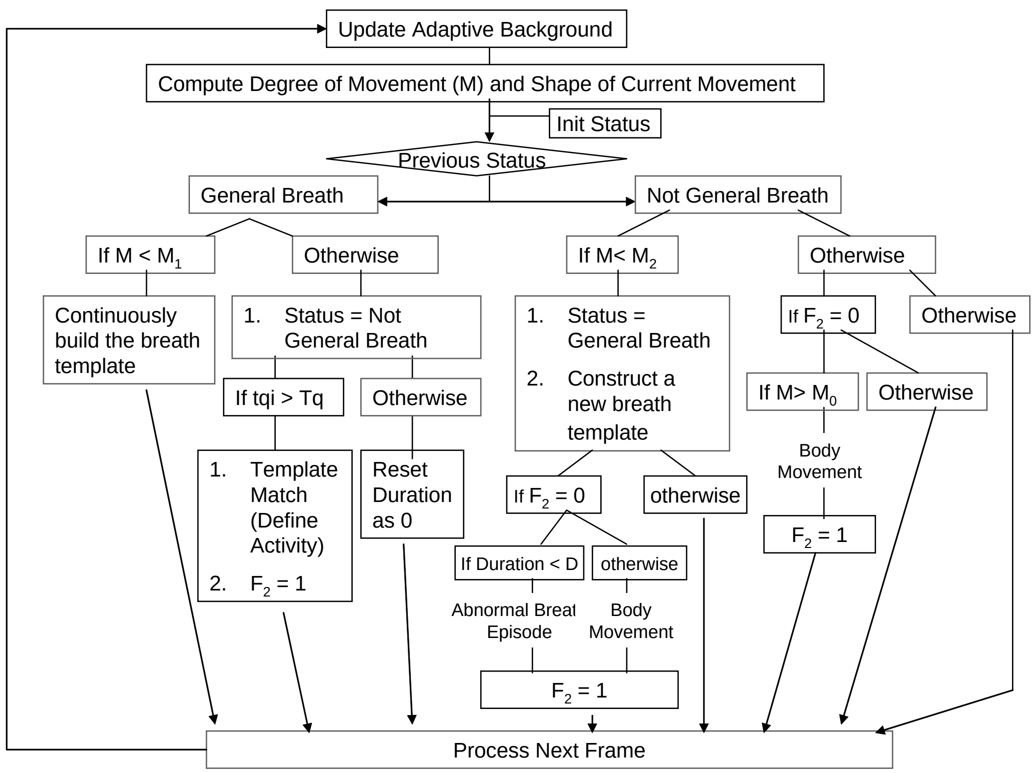


Figure 4. Experimental results

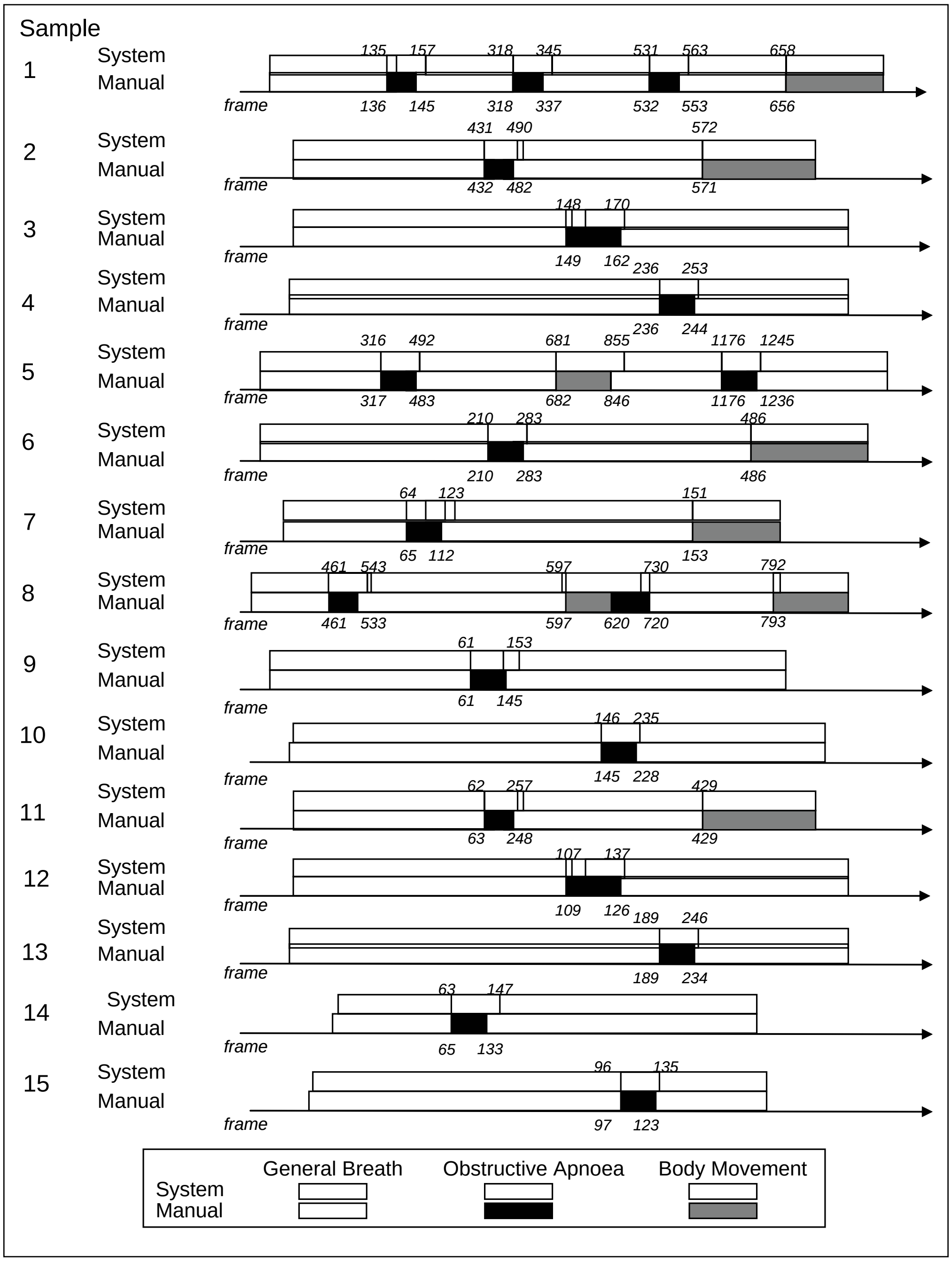
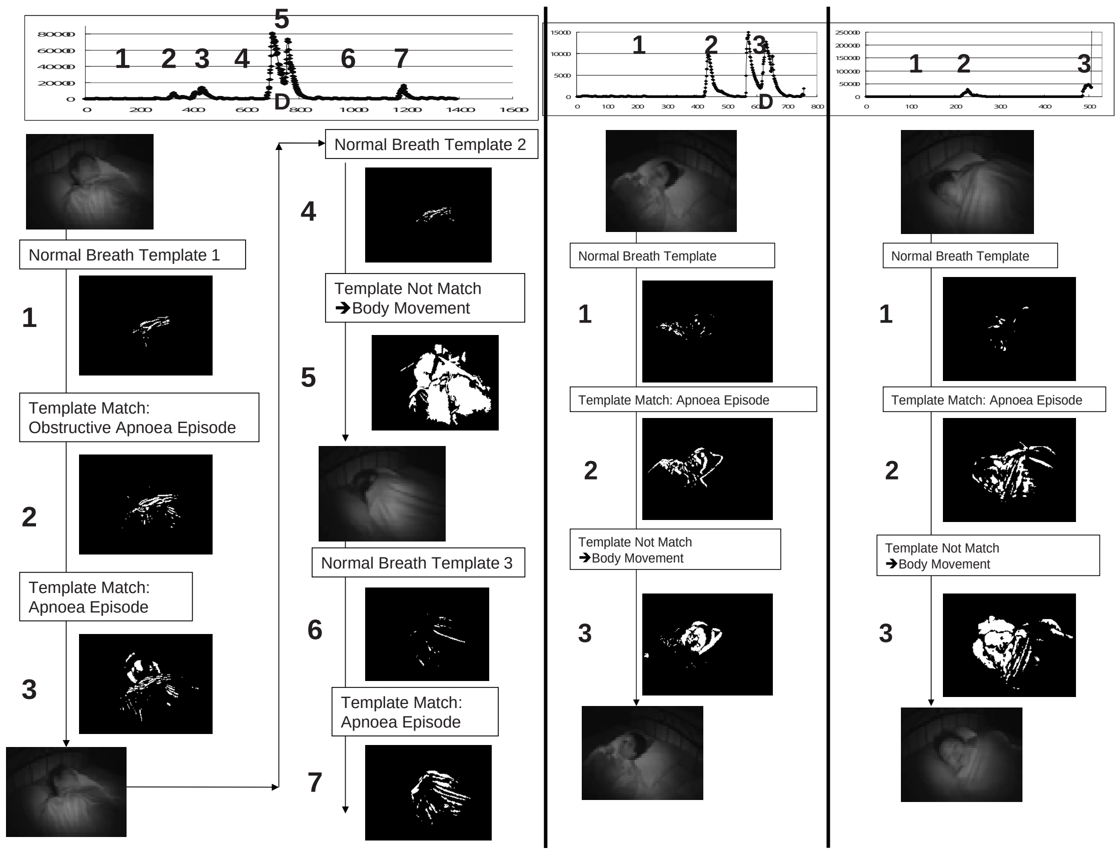


Figure 5. Illustration of three experimental analysis outputs



FUTURE TRENDS

We are investigating more automatic ways to obtain the threshold values and adapt them to individual subjects in order to accommodate human variations. More detailed experimental analysis using real rather than simulated episodes of Apnea will be conducted. We are also planning to support the video analysis of human breathing activity by adding audio analysis.

CONCLUSION

In this chapter, we present a noninvasive approach for recognizing abnormal breathing activity in assisting diagnosis of obstructive sleep apnea. The proposed approach utilizes infrared video information and avoids imposing geometric constraints or position limitations. The method is validated with limited experiments simulating the real environment for diagnosis of sleep disorders. Although the presented approach is mainly developed for diagnosis of obstructive sleep apnea, we believe that it can be utilized in other medical applications that require monitoring of the breath activity.

ACKNOWLEDGMENT

The PhD scholarship of Ching Wei Wang is jointly supported by the United Lincolnshire Hospital NHS Trust and University of Lincoln. We thank Dr. Neil Gravill and Dr. Simon Matusiewicz for their valuable comments in obstructive sleep apnea.

REFERENCES

Amin, H.M., Cigada, M., Fordyce, W.E., & Camporesi, E.M. (1993). Noninvasive monitoring of respiratory volume. Experimental evaluation of a breath monitoring device. *Anaesthesia*, *48*(7), 608–610.

Bennett, L.S., Langford, B.A., Stradling, J.R., & Davies, R.J.O. (1998). Sleep fragmentation indices as predictors of daytime sleepiness and nCPAP response in obstructive sleep apnea. *Am J Respir Crit Care Med, 158*, 778–786.

David, C.M., et al. (2003). Non-invasive analysis of physiological signals (NAPS): A vibration sensor that passively detects heart and respiration rates as part of a sensor suite for medical monitoring. In *Proceedings of the Summer Bioengineering Conference*, Key Biscayne, FL.

David W.G. (2005). Heart-respiratory monitor—infants. A.D.A.M., Inc. Retrieved May 5, 2006, from http://www.nlm.nih.gov/medlineplus/ency/article/007236.htm

Flemons, W.W., et al. (2003). Home diagnosis of sleep apnea: A systematic review of the literature: An evidence review cosponsored by the American Academy of Sleep Medicine. *American Thoracic Society CHEST, 124*(4), 1543–1579.

Hyun Medics. (2006). Baby breath monitor. Retrieved February 20, 2007, from http://hyun.en.ecplaza.net/14.asp

Kingshott, R.N., et al. (2000). Predictors of improvements in daytime function outcomes with CPAP therapy. *Am J Respir Crit Care Med, 161*, 866–871.

Murthy, R., Pavlids, I., & Tsiamyrtzis, P. (2004). Touchless monitoring of breathing function. In *Proceeding of the 26th Annual International Conference of the IEEE EMBS* (pp. 1196–1199), San Francisco.

Pepperell, J.C., Davies, R.J., & Stradling, J.R. (2002). Sleep studies for sleep apnoea. *Physiological Measurement, 23*, 39–74.

Randall, D.P. (1995). Remote respiratory monitor. In *Proceedings of the 8th Annual IEEE Symposium on Computer-Based Medical Systems* (pp. 204–211), Lubbock, TX.

Sivan, Y., Kornecki, A., & Schonfeld, T. (1996). Screening obstructive sleep apnoea syndrome by home videotape recording in children. *Eur Respir J, 9*, 2127–2131.

Storck, K., Karlsson, M., Ask, P., & Loyd, D. (1996). Heat transfer evaluation of the nasal thermistor technique. *IEEE Transactions on Biomedical Engineering, 43*(12), 1187–1191.

Svetlana, I., et al. (2005). A gated deep inspiration breath-hold radiation therapy technique using a linear position transducer. *Applied Clinical Medical Physics, 6*(1), 61–70.

Zhu, Z., Fei, J., & Pavlidis, I. (2005). *Tracking human breath in infrared imaging*. In *Proceeding of the 5th IEEE Symposium on Bioinformatics and Bioengineering* (pp. 227–231), Minneapolis, MN.

KEY TERMS

Continuous Positive Airway Pressure Therapy (CPAP): A therapy that delivers air into the patient's airway through a specially designed nasal mask or pillows. It is considered the most effective nonsurgical treatment for the alleviation of snoring and obstructive sleep apnea.

Hypopneas: Reductions in airflow or respiratory effort during sleep.

Infrared: Infrared light lies between the visible and microwave portions of the electromagnetic spectrum. It is used in night-vision equipment when there is insufficient visible light available.

Oxygen Saturation: A measure of how much oxygen the blood is carrying as a percentage of the maximum it could carry. It can be obtained from pulse oximetry.

Polysomnography (PSG): A diagnostic test in which a number of sensor leads are placed on the patient during sleep to record brain activity; eye, jaw muscle, and leg muscle movements; airflow; respiratory effort; heart rhythm; and oxygen saturation.

Sleep Apnea: A sleep disorder that causes breathing to stop during sleep for anywhere from 10 seconds to several minutes.

Thermal Imaging: An analogue pictorial representation or visualization of temperature differences.

APPENDIX

Pseudo-Codes

Inputs:
P*i* (*x*, *y*), the intensity value at location (*x*,*y*) at time *i*.
M0: Threshold to determine events as body movement
D: Threshold for time duration of the apnoea event
TQ: Threshold to use/not use the current template
M2: Threshold to switch from "not general breath" to "general breath" state
N: Noise filtering threshold for each individual pixel P*i*(*x*, *y*)
M3: Motion threshold for initializing S [step 4.1]
nt1: Number of frames with low motion degree for initializing S [step 4.2]
nt2: the number of the frame for template matching
nt3: the number of the frame for updating M1
t2, t3: threshold for template matching

Output:
Predicted result Ri at time i, where R = { 0: Apnea, 1: Body Movement, 2: Breath}
Parameter:
S: (0: not general breath, 1: general breath)
F1: (0: template match On, 1: template match Off)
F2: (0: define current event On, 1: define current event Off)
M1: Motion threshold to switch from "general breath" to "not general breath" state
B*i* (X,Y); the adaptive background at time *i*
T*i* (X,Y), the breath template at time *i*
S*i* (X,Y), the shape of current movement at time *i*
M: the degree of current movements
C1, C2, C3, C4, C5: counters
tqi: Template quality index (the number of data points) in the breath template

Steps:
1. Initialise background: B1(X,Y) = P1(X,Y)
2. Update the background: if i % 3= 0, then update the background (Adaptive Background Model)
   2.1 ΔI(*x, y*) = P*i* (*x, y*) – B*i-1*(*x, y*)
   2.2 if ΔI(*x, y*) > 0, then
      2.2.1 if(ΔI(*x, y*) < 1), then B*i* (*x, y*) = B*i-1*(*x, y*) + 1
      2.2.2 Else, B*i* (*x, y*) = B*i-1*(*x, y*) + ΔI(*x, y*)
   2.3 Else if ΔI(*x, y*) < 0, then
      2.3.1 if(|ΔI(*x, y*)| < 1), then B*i* (*x, y*) = B*i-1*(*x, y*) – 1
      2.3.2 Else, B*i*(*x, y*) = B*i-1*(*x, y*) – ΔI(*x, y*)
3.Compute Degree of Movement and Shape of current movement
   3.1 ΔII(*x, y*) = P*i* (*x, y*) – B*i* (*x, y*)
   3.2 if (ΔII(*x, y*) >= N), S*i* (*x, y*) = 255 and M = M + 1.
   3.3 Else, S*i* (*x, y*) = 0
4. Initialize S: if S is not initialized yet, then
   4.1 if(M< M3), C1=C1+1.
   4.2 if(C1> nt1), S = 1. Else, process next frame.
5. if F1 = 0, then
   5.1 C2 = C2 +1
   5.2 if(C2> nt2), then do template matching & define the event
      5.2.1
      ,where *n*(*x*,*y*)=1 if Ti(*x*,*y*) > 0 S*i* (*x*,*y*) > 0
      5.2.2
      ,where *n*(*x*,*y*)=1 if Ti(*x*,*y*) = 0 S*i* (*x*,*y*) > 0
      5.2.3 if(W2/ W1 > t2), then Ri = 1
      5.2.4 else if (W2/ W1 > t2), then Ri =0
      5.2.5 else, Ri=2
      5.2.6 F1 =1, F2= 1, C2 =0
      5.2.7 Process next frame
   5.3 Else, process next frame.

**if previous status is "general breath", check if need to change status or update the breath template
6. If Si-1 = 1, then
   6.1 If(M>M1), then
      6.1.1 S = 0
      6.1.2 If (tqi>TQ), F1 = 0
      6.1.3 Else, C3 = 0
      6.1.4 Process next frame
   6.2 Else, update breath template and define M1 based on the person's general breath motion degree.
      6.2.1 if T*i* (*x, y*) = 0 S*i* (*x, y*) > 0, then
         T*i* (*x, y*) = 255 and tqi = tqi +1
      6.2.2 C4 = C4 +1
      6.2.3 if (C4 =nt3), then
         6.2.3.1 M1 = tqi * 2.2.
         6.2.3.2 if(M1<400), M1 = 400

**if previous status is "not general breath", check if need to change status or define the event
7. Else, then
   7.1 C3 = C3 +1
   7.2 if (F2 =0 and M > M0), then Ri= 1 and F2 =1
   7.3 Else if (M < M2), then
      7.3.1 C5 = C5 +1
      7.3.2 if (F2 =0), then define the event
         7.3.2.1 F2 = 1
         7.3.2.2 if (C3 > D), Ri =1
         7.3.2.3 Else, Ri=0
      7.3.3 If(C5 > 10), then
         7.3.3.1 S = 1
         7.3.3.2 T*i* (*x, y*) = 0
         7.3.3.3 tqi =0, C4=0, F2=0, C5=0
   7.4 Process next frame

Detection of Gait Patterns in Challenging Environments

Kate Lynch
Victoria University, Australia

Daniel T. H. Lai
Melbourne University, Australia

Rezaul Begg
Victoria University, Australia

FALLS AND AGEING HEALTH CARE

One in three individuals over the age of 65 years (elderly) will fall at least once a year (James, 1993). This probability increases to one in two adults over 80 years (DHA, 2005). Consequently, gait modifications associated with ageing have been linked with increased falls' probability (Berg, Alessio, Mills, & Tong, 1997; Lord, Sherrington, & Menz, 2001). Despite an increasing research interest in recent times into the aetiology of falls, particularly in the elderly (>65 years), falls continue to be a major public health concern in Australia and worldwide. Fall-related injuries are the leading cause of accidental death in the elderly population, and account for the largest cause of hospitalisation for this population (Lord et al., 2001), with many elderly individuals experiencing physical, social, or functional ramifications following a fall. Consequently, the economic cost of falls to the public health system is escalating, with the total cost of fall injuries reported to be higher than road traffic injuries (DOH, 2004). The majority of falls associated costs include physician consultation, hospital stays, nursing homecare, rehabilitation, medical equipment, home modification and care, community based services, and prescription drugs and administration (DOH, 2004; Lord et al., 2001). Healthcare and related costs associated with falls are expected to double over the next 50 years (Close & Lord, 2006).

GAIT CHARACTERISTICS AND TRIPPING

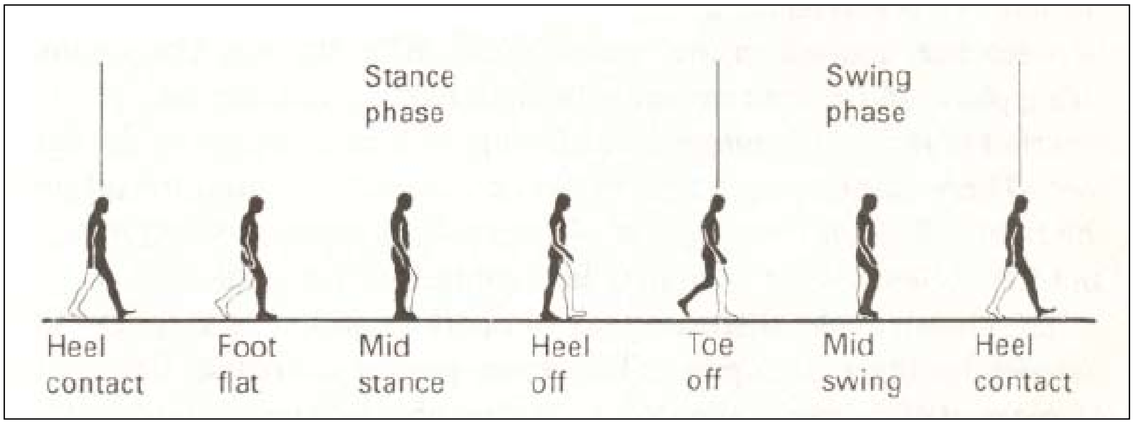
Walking or gait is the most common form of locomotion. A gait cycle is defined as the interval of time linking two successive occurrences of repetitive events of locomotion (most commonly foot contact/strike) (Kaufman & Sutherland, 2006; Whittle, 2003). The two main features of a gait cycle are standing and stepping, which are more technically known as stance and swing phases, respectively illustrated in Figure 1.

Typically the stance phase comprises approximately 60% of the gait cycle and is defined as the period when the foot is in contact with the ground. The swing phase makes up the remaining 40% of the cycle and is defined as the period between toe off to the second ipsilateral foot strike (i.e., when the foot is not in contact with the ground) (Kaufman & Sutherland, 2006; Russek, 1996; Winter, 1991). The purpose of the swing phase is to progress the body (most often in a forward motion), through sufficient clearance of the ground and to align the foot for the forthcoming heel contact. This is accomplished with the foot being swung through the air at a fast velocity and at a distance close to the ground (Perry, 1992; Winter, 1991).

Tripping refers to an unintentional disruption to the locomotion of the swing limb during a gait cycle, and consequently momentary suspension of the limbs' locomotion. This disruption can occur through an altercation with either the walking surface (ground) or an obstacle (e.g., mats, electrical cords) (AGC, 2002). Tripping is usually associated with two consequences: (i) stumble (in which balance is momentarily lost, however recovery occurs, and consequently does not result in a fall) or (ii) fall.

For safe uninterrupted trajectory during swing (to avoid tripping), the distance between the hip and the toe needs to be less than the distance between the hip and the ground surface or tripping hazard (Moosabhoy & Gard, 2006). The movements of the swing limb are a

Figure 1. The events of the gait cycle (Adapted from Whittle, 1991)



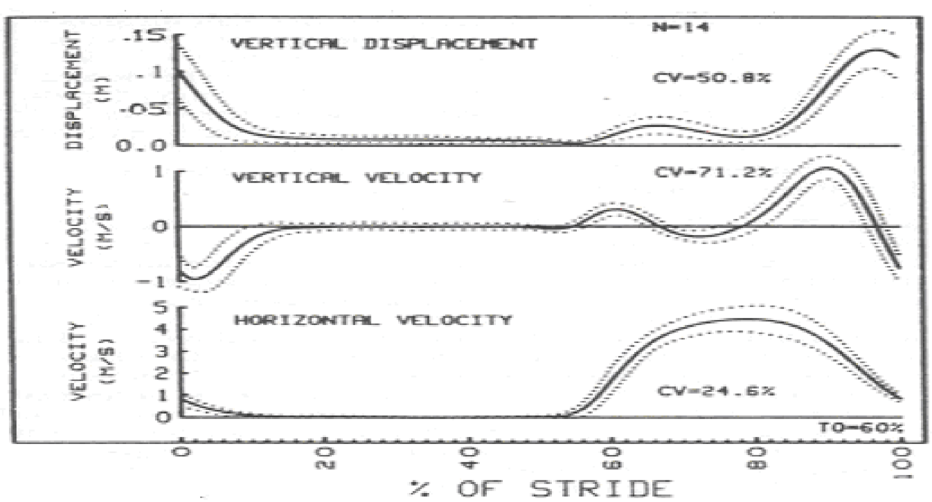
precise coordinated movement that is dependent upon a multisegment motor control chain of both stance and swing limbs (Winter, l991).

During the swing phase (60-100% of the gait cycle), the toe vertical trajectory undergoes two peaks as illustrated in Figure 2. Aforementioned, tripping avoidance is dependent upon adequate clearance of obstacles; it is the depression in vertical height between the two peaks in toe trajectory that is critical in order to avoid tripping (Weerdesteyn, Nienhuis, & Duysens, 2005; Winter, 1991). This drop in height is termed minimum toe clearance (MTC) and is a measurement between the lowest point of the foot (most commonly the toe) to the gradient of the walking surface during swing (Moosabhoy & Gard, 2006). The objective of MTC is to propel the limb forward with minimum energy expenditure (Perry, 1992). As a result of the timing, positioning, and velocity of the foot during low toe clearance, MTC height is the main contributing factor in tripping (Begg, Best, Dell'Oro, & Taylor, 2007; Murray & Clarkson, 1966; Winter, 1991).

Unfortunately gait modifications associated with ageing have been linked with an increased tripping risk and the ability to recover. Consequently, tripping is attributed as being the highest precipitating factor in elderly falls (35-40%) (Berg et al., 1997; Lord et al., 2001).

Prior investigations (Begg, Best, Dell'Oro, & Taylor, 2007; Winter, 1991) have found that the elderly demonstrate a lower MTC than the young and thus have an

Figure 2. Illustrates the mean vertical displacement and vertical and horizontal velocity for the period of one cycle of nine participants walking a natural cadence in Winter (1991)



increased risk of tripping. This combined with studies on ageing characteristics associated with an increased possibility of a trip becoming a fall (i.e., higher tripping frequency, delayed reaction time), provides a possible explanation for the high preponderance of falls in the elderly.

Winter (1991) calculated that at the precise moment of MTC height is controlled not only by the dorsiflexion/plantarflexion of the ankle, but is also equally influenced by four other motions, that is swing and stance knee flexion, stance ankle dorsiflexion/plantarflexion, and stance hip abduction/adduction.

ENVIRONMENTAL CIRCUMSTANCES

The control of MTC is not exclusively influenced by the events of the ankle joint however, but rather it is a pinpoint ballistic movement, controlled by lower limb joint kinematics of both the stance and swing limbs (Winter, 1991). Consequently, MTC can be affected and influenced by both intrinsic (internal) and extrinsic (external) modifications. Intrinsic factors that could possibly influence MTC height include changes associated with neurological diseases, deterioration in peripheral vision, and modifications associated with ageing. Extrinsic factors include shoes, visibility, and terrain characteristics (Butler, Druzin, & Sullivan, 2006; DOH, 2004). Typically a fall is a multifactorial phenomenon which occurs through a breakdown in an individual's ability (intrinsic) to successfully negotiate the environment (extrinsic).

GAIT MODIFICATIONS ASSOCIATED WITH AGING

Spatial-Temporal Gait Parameters

As previously mentioned, the two main features of the gait cycle are the stance and swing phases. During analysis of elderly and young adult groups on a flat surface, the elderly have been found to display longer stance duration than the younger adult group, and consequently a shorter swing phase. As a result of the shorter swing phase, the elderly group were also found to exhibit relatively longer double support duration (time when both feet are in contact with the ground) (Barela & Duarte, 2006; Mills & Barrett, 2001). It is possible however, that this increased stride duration is a result of the reduced walking velocity displayed by the elderly. For example, when Mills and Barrett (2001) accounted for velocity differences through normalization, the stance duration discrepancies were no longer significant. It is also plausible to suggest that the elderly have increased stride duration to allow a longer double support time to regain balance and control.

Lower Limb Joint Kinematics: Elderly

Barela and Durate (2006) analysed the differences in biomechanical characteristics between elderly individuals (age: 70 ± 6 years) and young adult individuals (age: 29 ± 6 years) whilst walking on a flat surface. Barela and Durate (2006) found that elderly individuals had a significantly smaller range of motion at the ankle joint than the young adult group. As a result, the elderly group had significantly less dorsiflexion than the young adults at the beginning of the swing phase. Research by Kerrigan, Lee, Collins, Riley, and Lipsitz (2001), however, revealed no significant differences between their elderly group (age: 73 ± 5.6 years) and their young group (age: 28.1 ± 4.2 years) in the ankle and knee joint angle kinematics. Kerrigan et al. (2001) did, however, find significant reductions in elderly peak hip extension in comparison to the young group, even when walking at a similar velocity. Due to this, elderly individuals with reduced peak hip extension will display reduced stride length.

GAIT MODIFICATIONS ASSOCIATED WITH SLOPED SURFACES

Spatial-Temporal Gait Parameters: Sloped Surfaces

Modifications to gait have been observed whilst walking on sloped surfaces; this is not surprising considering the constraints imposed on the individuals are different from those imposed on a flat surface. These include variations in gait in order to raise and lower the body's centre of mass to coordinate the vertical displacement of lower limbs during each stride and to ensure sufficient foot clearance (Leroux, Fung, & Barbeau, 2002; McIntosh, Beatty, Dwan, & Vickers, 2006). Although there is a substantial degree of literature dedicated to gait modifications in both the ageing and environmental situations, there is little, if any, literature dedicated to elderly individuals walking on sloped surfaces.

During analysis on eight healthy adult males (age 34 ± 9.4 years) who walked at a self selected pace on a treadmill with five inclines between -10% to 10%, Leroux et al. (2002) discovered modifications between all inclines for temporal-spatial data. Leroux et al. (2002) determined that during positive slope walking, there was a significant increase in stride length as the gradient angle increased. Leroux et al. (2002) also observed a progressive decrease in stride length as the gradient decreased from 0% to -10%. Research by McIntosh et al. (2006) indicated that during positive slope walking, there was a significant linear trend of decreasing cadence as the gradient angle increased from -10° to 0°. Cadence was also found to decrease from 0° to +10°, however, the change was not significant. Similarly a significant increase in speed occurred during this gradient modifications. Leroux et al. (2002), however, found that the proportion of stance/swing time, relative to cycle duration, was not affected by modifications in the slope

Lower Limb Joint Kinematics: Sloped Surfaces

Research focused on lower limb joint angle kinematics in slope walking discovered a number of significant modifications in comparison to level walking.

During analysis of upslope walking, research investigations found augmented increased flexion across the three lower limb joint angles (ankle, knee, and hip) from mid swing to early stance (Leroux et al., 2002) as the slope became progressively steeper. More specifically, during analysis on uphill walking Hansen, Childress, and Miff (2004) discovered the ankle to be the main adapting joint with the ankle significantly more dorsiflexed for the first 50% of the gait cycle to account for the increased gradient of the walking surface. The ankle is also dorsiflexed during the latter part of the swing to ensure the toe is raised higher than the ground surface in order to prevent a trip. The change at the ankle joint alone is not sufficient enough to ensure toe clearance of the raised gradient and the modifications to the trajectory of the swing limb are dependent upon modifications of both the stance and swing limb (Lay, Has, & Gregor, 2005)

Consequently, research conducted by Lay et al. (2005) found the knee to increase flexion at heel strike during upslope walking. However, increased extension was observed at mid stance in order to lift the body up the incline. Additionally, modifications of increased flexion and extension at the hip joint were also observed to ensure safe trajectory of the limb up the incline.

Modifications to the hip, knee, and ankle joint during down slope walking were (Leroux, Fung, Barbeua, 2002). However, the knee was determined to be the main adapting joint for downhill walking (Kuster et al., 1995; Leroux, 2002; McIntosh et al., 2006). Through increased knee flexion during stance, the length on the stance leg is shortened to allow for the swing leg to appropriately accept the ramp surface. According to Hansen et al. (2004), this is logical, as if only ankle plantarflexion was altered, the person would roll to the end of the stance foot and the swing limb would be inappropriately positioned to accept a level surface, resulting in an uncontrolled transition to the next foot, which would produce a tripping reaction.

The above findings indicate that the human locomotion pattern is highly adaptable to varying terrain (Leroux, 2002; Shkutatova, et al., 1999). These findings, however, were conducted on young healthy adult populations and consequently populations that lack range of motion at the hip, knee, or ankle joints (e.g., elderly, limited hip ROM), may have to compensate with an increased range of motion in their other joints, in order to sufficiently propel the body up and down the incline (McIntosh et al., in press).

MTC MODIFICATIONS ASSOCIATED WITH AGEING AND SLOPED SURFACES

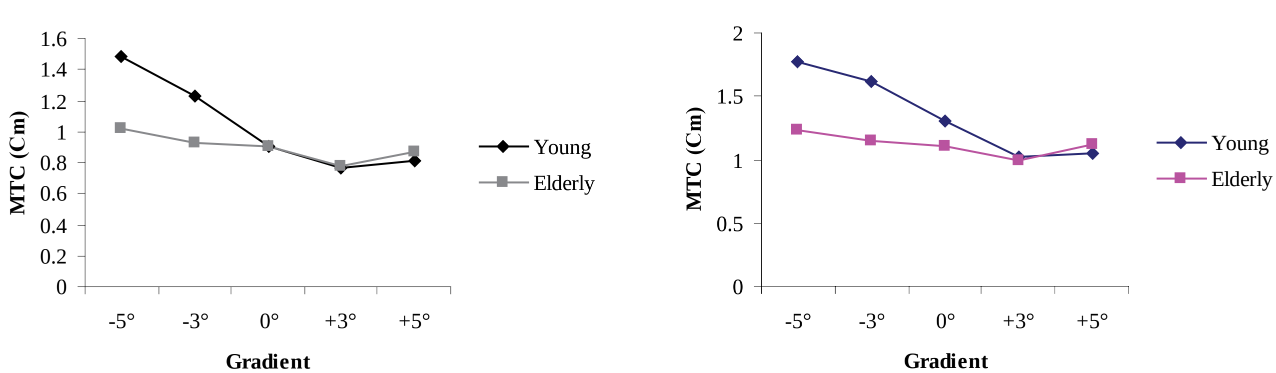
This following section is a comparative analysis of the modifications to MTC height for both elderly and young as a result of the challenges imposed by walking on a sloped surface. It is reasonable to hypothesise (given the results of previous sections) that there would be modifications to elderly MTC height as a result of both walking on a slope and also due to ageing effects. In order to test this hypothesis, we conducted our analysis on nine healthy young (18-35 years) and nine healthy elderly (65+) female adults. All participants walked at a self-selected speed at slopes of -5°, -3°, 0°, +3°, and +5° on a motorised treadmill. Note that each participant selected a speed that was consistent and walked across all gradients. Participants walked at each grade for approximately 7.5 minutes, in order to balance a sufficient number of cycles to obtain an accurate representation of the participants walking pattern without fatigue. Through the creation of imaginary markers (a point established through coordinates without the presence of a real marker) in Northern Digital Inc. (NDI) Toolbench, three imaginary markers were placed at the medial, frontal, and lateral locations at the front of the shoe, with the medial and lateral markers overlying the first and fifth-metatarsal phalangeal joint. An Optotrak Certus (an active-marker optoelectronic motion analysis system, NDI, Waterloo, Canada) was used to analysis gait kinematic data and through the use of a custom made Qbasic program, MTC height was calculated for each participant (Begg et al., 2007).

Limb movements of gait are asymmetrical (Taylor, 2003), and consequently in our study all participants' limbs displayed asymmetrical MTC heights, throughout all testing conditions. This indicates that one limb is travelling closer to the ground; hence one limb is at a higher risk of tripping. As tripping risk is largely influenced by MTC height, the events which occur on the low limb (section a of Figure 4) are more crucial in assessing tripping risk than the high limb. Hence, low limb characteristics will be the focal point of the following sections.

The negative gradients resulted in an increased MTC for both elderly and young groups. A repeated measure ANOVA found that the changes in MTC height between 0° and -3° (p=0.08) and 0° and -5° (p=0.15) gradients were not significant. The change, however between 0° and -5° was significant (p=0.02), suggesting there is a decreased probability of tripping at -5° gradient in comparison to a flat surface (0°). The modifications in the elderly group were neither as large nor as consistent with the modifications to gradient change in the young group. The young group increased MTC by 37% between 0° and -3° (0.90cm to 1.23cm). A further increase in MTC occurred between -3° and -5° (1.23cm to 1.48cm) of 20%, resulting in an overall increase of MTC height between 0° and -5° gradient of 63%. The modifications to MTC were considerably smaller in the elderly group. The elderly group only increased MTC by 2% between 0° and -3° gradients (0.91cm to 0.93cm), whilst despite the smaller gradient change the elderly group further increased MTC by 10% between -3° and -5° (0.93cm to 1.02cm), resulting in an overall MTC increase of 12% between 0° and -5°. This suggests perhaps ageing is associated with a reduced ability to appropriately alter gait to the constraints imposed by negative gradient modifications.

In contrast to walking on negative slope, a trend in MTC heights between positive slopes for both elderly and young was not apparent. Also inconsistent with the negative slope was the reduction in MTC height, in comparison to 0°. A repeated measures ANOVA indicated that groups significantly reduced MTC height between 0° and +3° (p=0.04). The height of both the young and elderly groups at +3° was 0.76cm and 0.78cm respectively. This lower MTC in itself suggests an increased likelihood of tripping at +3°, in contrast to the four other analysed gradients. The MTC height at +5° gradient was less than MTC at 0°, but larger than MTC at +3°, for both groups and limbs. However, no significant differences occurred between MTC values at 0° and +5°, and +3° and +5°. This suggest the toe is

Figure 4. Modifications to MTC height for elderly and young groups across all gradients: (a) Low limb, (b) high limb.



travelling closer to the ground at +5° in comparison to 0°, but higher from the ground at +5° in contrast to +3°. Similar to the other analysed gradients, the young group MTC height variations when compared to a flat surface 0° were larger than the elderly group (0.04cm).

Increasing MTC height results in higher energy expenditure (Begg, et al., 2007), whilst to successfully negotiate walking on raised surfaces increaed flexion and extension is required to elevate the limb above the raised suface (Prentice et al., 2004). This suggests positive gradients require higher energy expenditure than 0° gradients. This possibly serves as an explanation for the reduced MTC at positive gradients, as it suggests the body is unable to sufficiently match the mechanical demands for walking up a sloped surface. In this research study design, participants were required to walk for a considerably long period of time at the elevated gradient (approximately 7.5 minutes), which may have had an effect on the extent of increased energy expenditure. If however the reduced MTC is a consequence of the demands of the raised surface, it serves as no explanation for the lower MTC at +3°, in comparison to +5°. One suggestion, however, is that +3° is not a large enough modification to gradient for participants to consciously adjust gait, and as a result the limb travels closer to the ground.

COMPUTATIONAL INTELLIGENCE IN GAIT ANALYSIS

Computational intelligence (CI) in gait analysis generally refers to the fusion of machine learning paradigms and computer science resulting in powerful classifiers which learn data relationships. Popular CI techniques have focused on supervised learning and include artificial neural networks (Haykin, 1994), support vector machines (Vapnik, 2000), neurofuzzy classifiers (Abe, 1997), and hybrid models. In supervised learning, a classifier learns the implicit relationship between input data and the predefined class labels. This relationship may be a function describing a complex biomechanical system; for example, balance control in the lower extremities during locomotion. While prior linear discriminant analysis investigated distribution characteristics of particular gait variables, newer CI techniques offer more powerful nonlinear analysis methods.

CI techniques have only recently been applied to research in human movement sciences and gait analysis when it was realised that locomotion was vital in our daily activities and our perceived quality of life (Whittle, 2003). Numerous gait variables such as joint angles (Oberg, Karsznia, & Oberg, 1994) and ground reaction forces (Begg, Sparrow, & Lythgo, 1998) can be recorded making feature selection an important aspect of gait data classification. CI has been recently used to study normal gait biomechanics (Barton & Lees, 1997; Barton, Lisboa, Lees, & Attfield, 2007) and detect gait pathologies arising from lower extremity injuries and imbalance disorders due to cerebral diseases such as Parkinson's (Morris, Iansek, Matyas, & Summers, 1994) and Cerebral Palsy (Kamruzzaman & Begg, 2006).

Falls in the elderly are another major health concern since they cause serious injuries and can be fatal, hence motivating the need for newer prediction and prevention techniques (National Safety Council, 2005). Recently CI techniques have been used to detect and predict elderly persons at risk of falling using gait variables such as the minimum toe clearance (Begg, Lai, Taylor, & Palaniswami, 2005; Khandoker, Lai, Begg, & Palaniswami, 2006).

Example: Classification of Elderly and Young from Inclined Walking

We present a classification example to examine the issues involved in the application of CI to gait problems. In this example, we hypothesize that minimum toe clearance (MTC) measurements could better discriminate healthy elderly from young gait if the subjects were made to walk on different ground inclinations. Statistical parameters of the MTC data were computed and used as input features to train a Support Vector Machine (SVM) classifier.

Altogether, seven statistical features were computed from the MTC data distribution which included the mean, standard deviation, skewness, kurtosis, median, first quartile (Q1), and third quartile (Q3) values. The SVM was trained using the linear, polynomial, and Gaussian kernel over a range of $C=\{0.1,1,10,100,1000\}$. The best accuracy results are reported in Table 1 and suggest wide variation in success rates offered by the classifiers across the different inclinations. The results also indicate that the SVM is able to better differentiate the two age groups with an accuracy of 82.35% using MTC characteristics at positive 3º inclination. This accuracy rate, although moderate, is higher than those detected using MTC characteristics from flat

Table 1. Best leave one out SVM classification accuracies for various inclinations. Flat –flat or 0º walking surface, p3- positive 3º, p5- positive 5º, n3- negative 3º, n5- negative 5º

		LOW		
p5	p3	Flat	n3	n5
52.94	58.82	64.71	64.71	64.71

		HIGH		
p5	p3	Flat	n3	n5
58.82	82.35	47.06	52.94	76.47

(0º) walking condition. These results were obtained using a small sample size that might have affected the outcomes and, therefore, this need to be checked using a larger sample size and also using various other inclination angles in future studies. Nevertheless, these results are encouraging in the future applications of computational intelligence approaches for modelling gait data for the detection of ageing effects on gait control and falls risks.

FURTHER WORK

The application of CI techniques in gait analysis has not kept pace with other biomedical areas and further research is required to make them routine in clinical and rehabilitative settings. In particular, fuzzy and genetic algorithm type techniques have yet to be investigated in detail for gait problems. Further research in this area using various signal processing techniques and CI techniques is required to investigate the accuracy of detecting changes in gait due to negotiation of challenging environments.

REFERENCES

Abe, S. (1997). *Neural networks and fuzzy systems: Theory and applications*. Boston: Kluwer Academic.

AGC. (2002). *Guide to preventing slips, trips and falls*. Australian Government Comcare, Safety Rehabilitation and Compensation Commission. Canberra, Australia: ACT.

Barela, A., & Duarte, M. (in press). Biomechanical characteristics of elderly individuals walking on land and in water. *Journal of Electromyography and Kinesiology.*

Barton, J., & Lees, A. (1997). An application of neural networks for distinguishing gait patterns on the basis of hip-knee joint angle diagrams. *Gait and Posture, 5*, 28-33.

Barton, G., Lisboa, P., Lees, A., & Attfield, S. (2007). Gait quality assessment using self-organising artificial neural networks. *Gait and Posture, 25*, 374-379.

Begg, R., Best, R., Dell'Oro, L., & Taylor, S. (2007). Minimum foot clearance during walking: Strategies for the minimisation of trip-related falls. *Gait and Posture, 25*, 191-198.

Begg, R., Lai, D., Taylor, S., & Palaniswami, M. (2005). SVM models in the diagnosis of balance impairments. In *3rd International Conference on Intelligent Sensing and Information Processing*, Bangalore, India (pp. 248-253).

Begg, R., Sparrow, W., & Lythgo, N. (1998). Time-domain analysis of foot ground reaction forces in negotiating obstacles. *Gait and Posture, 7*, 99-109.

Berg W., Alessio, H., Mills, E., & Tong, C. (1997). Circumstances and the consequences of fall in the independent community-dwelling older adults. *Age and AGEING, 26*, 261- 268

Bulter, E., Druzin, M., & Sullivan, E. (2006). Gait adaptations in adulthood: Pregnancy, AGEING and alcoholism. In J. Rose & G. Gamble (Eds.), *Human walking* (3rd ed.). Sydney, Australia: Lippincott Williams and Wilkins.

Close, J., & Lord, S. (2006). *Falls in the elderly.* Retrieved January 31, 2008, from http://www.australiandoctor.com.au

DHA. (2005). Department of Health and Ageing. Australia Government. *The National Slops and Falls Prevention Project*. Retrieved January 11, 2006, from http://www.slope.schise.com/docs/uploads/HACC

DOH. (2004). *The falls policy for older western Australians*. Department of Health Government of Western Australia. Retrieved January 31, 2008, from http://www.stayonyoungfeet.com.au/documentation/falls_policy_for_older_wa

Hansen, A., Childress, D., & Miff, S. (2004). Roll-over characteristics of human walking on inclined surfaces. *Human Movement Science, 23,* 807-821

Haykin, S. (1994). *Neural networks: A comprehensive foundation*. Prentice Hall.

James, M. (1993). *Crime prevention for older Australians.* Canberra: Australia: Australian Institute of Criminology.

Kamruzzaman, J., & Begg, R. (2006). Support vector machines and other pattern recognition approaches to diagnosis of cerebral palsy gait. *IEEE Transactions on Biomedical Engineering, 52,* 828-838.

Kaufman, K., & Sutherland, D. (2006). Kinematics of normal walking. In J. Rose and G. Gamble (Eds.), *Human walking* (3rd ed.), Sydney, Australia*:* Lippincott Williams and Wilkins.

Kerrigan, D., Lee, L., Collins, J., Riley, R., & Lipsitz, L. (2001). Reduced hip extension during salking: Healthy elderly and fallers versus young adults. *Arch Phys Med Rehabil, 82,* 26-30.

Khandoker, A., Lai, D., Begg, R., & Palaniswami, M. (2006). A wavelet based approach for screening falls risk in the elderly using support vector machines. In *IEEE Proceedings of Fourth International Conference on Intelligent Sensing and Information Processing (ICISIP)* (Vol., pp. 78-82).

Kuster, M., Sakurai, S., & Wood, D.A. (1995). Kinematic and kinetic comparison of downhill walking and level walking. *Clinical Biomehcanic, 10*(2), 79-84

Lay, A., Hass, C., & Gregor, R. (in press). The effects of sloped surfaces on locomotion: A kinematic and kinetic analysis. *Journal of Biomechanics*.

Leroux, A., Fung, J., & Barbeau, H. (2002). Postural adaptation to walking on inclined surfaces: I normal strategies. *Gait and Posture, 15*(1), 64-74

Lord, S., Sherrington C., & Menz, H. (2001). *Falls in older people: Risk factors and strategies for prevention.* UK: Cambridge University Press.

McIntosh, A., Beatty, K., Dwan, L., & Vickers, D. (2006). Gait dynamics on an inclined walkway. *Journal of Biomechanics, 39,* 2491-2502.

Mills, P., & Barrett, R. (2001). Swing phase mechanics of the healthy young and elderly men. *Human Movement Science, 20,* 427-446.

Moosabhoy, M., & Gard, S. (in press). Methodology for determining the sensitivity of swing leg toe clearance and leg length to swing leg joint angles during gait. *Gait and Posture*.

Morris, M., Iansek, R., Matyas, T., & Summers, J. (1994). The pathogenesis of gait hypokinesia in Parkinson's disease. *Brain, 117*(5), 1169-1181.

Murray, C., & Clarkson, B. (1966). The vertical pathway of the foot during level walking: I range of variability in normal men. *Journal of the American Physical Therapy Association, 46*(6), 585-589.

National Safety Council. (2005). *Falls among older adults: A national problem*. White Paper for White House Conference on Ageing.

Oberg, T., Karsznia, A., & Oberg, K. (1994). Joint angle parameters in gait: Reference data for normal subjects, 10-79 years of age. *Journal of Rehabilitation Research and Development, 31,* 199-213.

Perry, J. (1992). *Gait analysis: Normal and pathological function.* New York*:* McGraw-Hill Inc.

Prentice, S. D., Hasler, E. N., Groves J., Frank., J. S. (2004). Locomotor adaptations for changes in the slope of the walking surface. *Gait & Posture, 20*, 255-265.

Russek, L. (1996). Closed kinetic chain and gait. In R. Donatelli, R. (Ed.), *The biomechanics of the foot and ankle.* Philadelphia, PA: Davis Company.

Vapnik, V.N. (2000). *The nature of statistical learning theory* (2nd ed.). New York: Springer.

Weerdesteyn, V., Niehuis, B., & Duysens, J. (2005). Advancing age progressively affects obstacle avoidance skills in the elderly. *Human Movement Science, 24*(5-6), 865-880.

Whittle, M. (1991). *Gait analysis: An introduction*. Oxford, England: Butterworth Hinemann.

Whittle, M. (2003). *Gait analysis: An introduction (3rd ed.*). Oxford, England: Butterworth Hinemann.

Winter, D. (1991). *The biomechanics and motor control of human gait: Normal, elderly and pathological* (2nd ed.). Ontario, Canada: University of Waterloo Press.

KEY TERMS

Computational Intelligence: The fusion of machine learning paradigms and computer science resulting in power classifiers which learn data relationships.

Elderly: Adults over the age of 65 years.

Falls: To leave a standing or erect position suddenly, whether voluntarily or not, and succumb to either a sitting, kneeling, or lying position.

Gait: A particular way or manner of walking.

Minimum Toe Clearance: The measurement between the lowest point of the foot (most commonly toe) to the gradient of the walking surface.

Stance: The period of time when the foot is in contact with the ground during gait.

Swing: The period of time between toe off to the strike of the same foot during gait.

Developing Medical IT Systems that Save Lives and Significantly Reduce Hospital Healthcare Costs

Robert J. Mockler
Tobin College of Business, USA

Dorothy G. Dologite
Baruch College, USA

INTRODUCTION

This chapter originated with strategic management work done at Jamaica Hospital in Queens, New York, and the Jewish Home and Hospital for the Aged in Manhattan, New York. As background for the project, the initial phase involved industrywide studies of health care institutions throughout the United States and abroad. During these studies, which involved both field research and a review of the research literature, many samples of which are given throughout this chapter, it became apparent that advances in hospital information technology (IT) are having a dramatic impact on improving patient health care services.

BACKGROUND

By creating the "Digital Hospital," a number of hospitals in this country are leading the way in developing and using information systems to solve plaguing industry problems, including the fact that manually kept records were not up-to-date, were incomplete, were misplaced, or were illegible due to physicians' handwritten prescriptions (Clark, 2004; Information Technology, 2005a, 2005b). Errors resulting from these problems alone have been estimated to kill more than 7,000 U.S. hospital patients yearly and drive up health care costs by an estimated $2 billion (Evans, 2004; Marks, 2004; Turner, 2004). Even though major costs are involved, plans are underway to do the same in overseas hospitals; for example, in England. Evidently, in 2004, not enough U.S. hospitals were moving into the digital area. As a result, in mid-2004, Tommy Thompson, Secretary of Health and Human Services, supported by President Bush, unveiled an ambitious 10-year initiative to transform hospitals in this country more aggressively and rapidly (Bush, 2004; Turner, 2004). Individual states (e.g., Massachusetts) were also considering both government and private financial support for this effort (Peter, 2004). The focus of this initiative will be on computer entry of prescriptions, improved intensive-care unit (ICU) staffing, and easier access to individual patient records for both the patient and hospital staff by integrating patient and other hospital records into integrated databases.

ROLE OF ADVANCED AND BASIC INFORMATION TECHNOLOGY APPLICATIONS AS IMPROVED HOSPITAL HEALTH CARE ENABLERS: A FRAMEWORK

The discussion in the following section provides an overview framework developed from this study, a framework that might help provide guidance to those thinking about understanding, developing, and introducing IT systems into their hospital environments. This framework is outlined in Figure 1.

CLIENT/PATIENT ORIENTATION

Hospitals face specific problems, such as the difficulty to coordinate and integrate diverse customer interactions in a way that (a) makes individual customers feel they were being treated as individuals and (b) enables the hospital to coordinate its services to these individuals. Hospitals have many divisions, such as radiology, emergency room care, test labs, prescriptions, medical nursing assistance, hospital rooms, and kitchens and

Figure 1. Major areas in which medical IT systems are generally developed

Client/Patient Orientation
An Integrated Accessible Electronic Database
Prescription Writing and Delivery
Intensive Care Units (ICUs)
Telemedicine/Telesurgery
Health Care Facilities and Their Financing
Other IT-Related Systems

food service, that provide services to patients. The doctors deal with these various divisions as well as with individual patients, as do nurses and other staff members whose time needs to be scheduled and managed. Tracking is needed to bill patients, preferably in a coordinated way; to maintain staff schedules; and to manage all operational areas efficiently and effectively. Interestingly, many hospitals studied do not even have an integrated billing system for patients, who still receive separate bills for various department services and room accommodations. This creates some confusion among patients and their insurance companies.

AN INTEGRATED ACCESSIBLE ELECTRONIC DATABASE

Based on hospital experiences with information technology development, the starting point almost always involves building accessible integrated electronic databases, especially in relation to individual patient information, as suggested in the preceding section. The most significant applications of such databases are in the prescription writing and delivery area and in the Intensive Care Unit (ICU) area. These findings are consistent with the authors' studies of individual hospitals and the areas that might benefit most from improvements.

Prescription Writing and Delivery

Probably the most visible evidence of the new digital systems at the Children's Hospital in Pittsburgh is the trolley used by physicians when visiting individual patients. The trolley has a laptop computer and wireless network card, which physicians use to log into a program containing all the clinical information on each of their patients. The physicians fill out medication orders online and send these orders from the ward on the seventh floor to the in-house pharmacy in the basement, where a robot puts the prescribed drugs in an envelope for the nurses to dispense. All of this is made possible by the electronic patient database. This laptop-enabled system significantly reduced the major problems encountered from medication errors under the old paper-based system. While medical information technology applications do not necessarily eliminate all errors (in fact, they introduce some new ones), they do substantially reduce these errors (Conn, 2005; Marks, 2004).

INTENSIVE CARE UNITS (ICUS)

Another major benefit of medical IT systems is found in the management and running of ICUs. For example, in late 2004, intensive care patients at Inova Alexandria Hospital were scheduled to be monitored around the clock by doctors and nurses at their bedside as well as through digital cameras, microphones, and special software from miles away (Salmon, 2004). This system will enable these distant caregivers to remotely monitor heart rates, blood pressure, respiratory rates, and other vital signs of critically ill patients even more closely than the on-site duty staff can and thus provide guidance to that duty staff. An IT-based ICU system, where one doctor and nurse can keep a 24-hour watch over as many as 200 critically ill patients at once, can boost chronically short-staffed on-site care. In addition, studies have reported as much as a 50% drop in ICU mortality and 17% shorter stays since the first such system was set up at Virginia's Sentara Health care a few years ago (Allen, 2004; Salmon, 2004).

TELEMEDICINE/TELESURGERY

One early IT development area has been telemedicine. For example, the U.S. military in Iraq set up a field hospital unit connected by computer systems to well-staffed hospital units located elsewhere. This new technology enables treating injured people in remote locations using local staff guided by expert professionals located elsewhere (Hasson, 2004). Kaplolani Medical Center for Women & Children in Hawaii has begun broadcasting live surgeries to physicians around the world. This telesurgery technology allows specialists worldwide to observe and consult with Kapiolani surgeons during emergency and other procedures. It also involves not having to move critically ill patients needing surgery, which is a major cost reduction factor and can help avoid additional medical problems caused by traveling. In addition, it enables an auditorium of medical students to observe live surgeries.

HEALTH CARE FACILITIES AND THEIR FINANCING

Hospitals need sufficient finances to obtain and operate these new facilities as well as to hire and maintain expert professional staff at all levels and put into place the IT digital framework to service everyone. A variety of financial factors affect a hospital's ability to do this. For example, a hospital is dependent on insurance company payments for a substantial portion of their income. A patient's ability to pay also has an impact; and there are limits to the number of lawsuits a hospital can pursue to recover delinquent accounts, and such lawsuits are time consuming and expensive. In addition, many communities expect or require hospitals to provide services for the poor.

IT solutions can be very helpful in the financial area. They enable more accurate billing of patients, faster collection of accounts payable, and easier coordination with insurance companies and government agencies, as well as better tracking of and control of staff usage (Paulik, 2004). In a very significant way, it has led to major reductions in cost-per-patient services, enabled improved health care service delivery, and reduced unnecessary loss of life.

OTHER IT-RELATED SYSTEMS

Integrated IT systems enable other individual patient health care service improvements. The following happened at Allegany General Hospital in Pittsburgh: nurse Candice Bena thought a 76-year-old patient needed a new intravenous line but couldn't get the radiology department to install one immediately (Wysocki, 2004). Fearing the patient would develop an infection, the nurse eventually contacted the Chairman of Medicine directly. The chairman immediately contacted the head of the Radiology Department electronically, who within two hours installed the new IV line himself. The result of such problem identification and process analysis was a 90% drop in the number of infections after just 90 days of using the system. As an additional bonus, the new system saved almost $500,000 a year in ICU costs.

THE BENEFITS CAN BE SUBSTANTIAL AND IMPRESSIVE

In addition to these benefits already discussed, Cincinnati Children's Hospital Center, for example, won the 2003 Nicholas E. Davies EMR (Electronic Medical Record) Recognition Award of Excellence after implementing an integrated clinical information system (ICIS) to provide clinical decision support tools, including Siemens Medical Solutions (INVISION®). The results were elimination of transcription errors, a 50% reduction in medication errors, a 52% improvement in medication turnaround times, a 24% reduction in verbal orders for controlled substances, and 100% compliance with pain assessment documentation requirements defined by state regulatory agencies (Siemens, 2005).

Delnor Community Hospital also showed impressive results with HeartMath's customized technology Freeze-Framer®. After the first year, the overall employee turnover was reduced from 28% to 20.9%, which led to $800,000 in annual savings. Medicare length of stay decreased by 9%, equaling a $1.4 million savings annually, customer satisfaction improved from the 73rd percentile to the 93rd percentile, and the hospital was ranked first in employee satisfaction based on a Sperduto and Associates study. During the second year, the results were maintained with turnover down to 14%, and the hospital was ranked second in employee satisfaction (HeartMath, 2005).

Other benefits included reducing patient waiting times; slashing wheelchair inventories; preparing operating rooms faster; and moving patients through a hospital stay or doctor visit more quickly, seamlessly, and error free, creating major cost reductions along the way in addition to improving patient health care, and in many instances actually reducing the death rate.

IMPLEMENTATION PROBLEMS

Although substantial gains from introducing and using medical IT systems are documented, major problems have been encountered. For example, outside physicians at many hospitals have resisted and in some instances even scuttled new medical IT systems (Turner, 2004). It is far easier to use a pen to write a prescription order or tell a nurse what to do than it is to have to learn a whole new system, no matter how fast or easy it may eventually be to use. The training needed to ease entry of the new systems takes many forms, depending on the staff involved. As Rob Turner points out, patience and persistence is the primary requirement, combined with a keen sense of individual needs in specific hospital situations (Turner, 2004). For example, at a university hospital where most of the physicians are on staff and thus are more controlled by hospital management, the transition has been easier than at hospitals where the majority of physicians are outsiders and thus more independent.

There are also group-training tools that can aid in implementation (Bandarouk & Sikkel, 2005). As might be expected, for example, the various characteristics of users centralized within an existing medical department (e.g., ICU) and those of distant (decentralized) users of any new system had a major impact on the managerial support tools needed to make the implementation successful. The successful management support tools differed in several areas: time allocated (more for the decentralized group); feedback and learning opportunities (different channels and leadership styles used for each group), and autonomy and responsibility (differs by individual as well as group).

A second key problem area is the high cost of such systems. The essence of making such systems work is the integrated computer systems and databases that provide the data/information and facilitate the processing of it. This veritable digital hub involves a substantial financial investment to develop. Again, in spite of the documentable savings, time is needed to study the application of other hospitals' experiences to one's own hospital. And once an appropriate transformation plan is developed, raising the money and convincing others of the investment's value is still a long hard road to travel (Bandarouk & Sikkel, 2005; Turner, 2004).

Given the varying situation-specific needs and financial constraints, effective planning and implementation require identifying individual situation requirements in conjunction with determining the specific components of the planned medical IT system, as well as the order in which the IT transformation steps will take, based for example on analyses similar to the ones given in the previous framework (Figure 1). This is, in a sense, defining the "migration path" to be followed and the basic management principles, from planning and leadership to measurement, guiding this migration (Amatayakul & Lazarus, 2004; Skinner, 2003).

FUTURE TRENDS

The growing interest in medical IT systems is not limited to the United States. In the United Kingdom, the National Health Service (NHS) signed 10-year contracts worth close to $10 billion in 2004 to provide electronic patient records, e-booking, e-prescribing, and electronic ordering of tests in arguably the world's biggest civil IT procurement—one that will cover all 50 million patients in England and one million staff. The program should also see digital images replace X-ray film, making remote consultation about a patient's condition possible. Even if all goes smoothly, it will take until 2008 at the earliest for the system to be operational. But the first NHS hospital to become virtually paperless—University College Hospital in central London—should complete this transformation in 2006. Other examples of countries that have been successful in implementing IT systems in hospitals are Germany, Sweden, and Thailand (Becker, 2004). Several U.S. hospitals have already made this transition, and many, if not most, are expected to follow once financing has been arranged. What is described here, then, is an international phenomenon that will revolutionize hospital medical delivery practices worldwide, which in turn will affect the health care industry in the United States.

Since predictions were that IT spending in health care would reach $47.5 billion by the end of 2006, many firms are entering the field. General Electric Medical Systems and Siemens Medical Systems (INVISION®) are two major corporations that have medical IT development divisions. Computer giants such as IBM, SAP, Hewlett-Packard, and Dell, as well as niche software specialists such as Cerner of Kansas; VISICU of Baltimore; Eclipsis of Florida; HeartMath of Boulder Creek, California; MercuryMD; Infocrosing, Inc.; and the merged Picis and Ibex, Inc. have also entered the field to gain a foothold in what is expected to be a very large market (Clark, 2004; Information Technology, 2005a, 2005b, 2004, 2005).

CONCLUSION

While it is important to review industry practices, the ultimate decision then is a situational one. Individual hospitals have their own special needs. For example, Jamaica Hospital and the Jewish Home and Hospital were each very distinct in the services they provided and the kind of clients they served; as a result, the IT systems they developed were different in many ways. Kapolani Medical Center in Hawaii was uniquely situated to develop its telesurgery program. At university hospitals, their close relationship with their physicians on staff on occasion enabled them to move more quickly into automated prescription processing.

External circumstances can also have an impact, as seen in Massachusetts where the state government support was expected to encourage and facilitate medical IT systems development.

In almost all instances, available finances will have an impact on the nature and pace of medical IT systems development, as will the leadership skills of hospital managers in overcoming resistance to change encountered at almost all operational levels.

In light of these problems, currently less than 2% of all American hospitals have fully implemented electronic medical records. At this time, however, 40 hospitals have portions of such a system and many more are considering introducing such systems. This slow development pace is in spite of the fact that experience shows that these systems can save lives and dramatically affect improvement of health care services delivery at hospitals. Hopefully, studies such as this one, as well as government programs and incentives, will help broaden and accelerate the growing trend toward continuing medical IT systems development worldwide.

REFERENCES

Allen, L. (2004). Warming up to telemedicine. Bleep! Blip! Ding! Another life saved. Why electronic intensive care isn't as scary as it seems. *Popular Science*. June, 64.

Amatayakul, M., & Lazarus, S.S. (2004). Aiming for more than a base HIT. *Healthcare Financial Management*. November, 98–101.

Bandarouk, T., & Sikkel, K. (2005). Explaining IT implementation through group learning. *Information Resources Management Journal*. January-March, 42–61.

Becker, C. (2004). The best care money can buy? *Modern Healthcare*. August 9, 26–30.

Bush, G.W. (2004). Transforming health care: The president's health information technology plan. State of the Union Address. January 20.

Clark, T. (2004). Hospitals offer prescription for profit: Healthcare. *Financial Times*. February 11, 13.

Conn, J. (2005). Dealing with errors. *Modern Healthcare*. January 3, 12–13.

Evans, B. (2004). Our nation's health depends on IT. *InformationWeek*, September 27, 82.

Hasson, J. (2004). U.S. military deploys new telemedicine tools in Iraq. *Federal Computer Week*. November 18.

HeartMath. (2005). Delnor-Community Hospital soars with HeartMath. Retrieved January 28, 2005, from http://www.heartmath.com/executive/case_study_health_care.html

Information technology: Atlanta hospital provides physicians with wireless access to patient data. (2005a). *Medical Devices & Surgical Technology Week*. January 9, 155.

Information technology: Network video technology provided for improved ICU patient care. (2004). *Obesity, Fitness & Wellness Week*. December 25, 830.

Information technology: Two VA hospitals to adopt patient progress software. (2005b). *Science Letter*. January 11.

Marks, A. (2004). The steps still to be taken in reducing medical errors; Healthcare experts call for more funding for high-tech upgrades, better training, and changes in institutional culture. *The Christian Science Monitor*, November 26, 2.

Mergers and acquisitions: Segment of Verizon information technologies healthcare business acquired. (2004). *Obesity, Fitness & Wellness Week*. October 2, 854.

Paulik, D. (2004). Breaking up the transcription logjam can improve cash flow. *Healthcare Financial Management*, June, 78–81.

Peter, J. (2004). Med files may end paper trail: State may be nation's first to have electronic records. *Telegram & Gazette*. December 7, A2.

Salmon, J.L. (2004). Alexandria Hospital to create a virtual ICU. *The Washington Post*. September 18, B2.

Siemens (2005). Cincinnati Children's Hospital Medical Center becomes first children's hospital to win Davies Award. Retrieved January 28, 2005, from http://www.medical.siemens.com

Skinner, R. (2003). The value of information technology in healthcare. *Frontiers of Health Services Management*. Spring, 3–16.

Turner, R. (2004). A high dose of tech: Some grocery stores have better technology than our hospitals and clinics. Tommy Thompson, Secretary of Health and Human Services. *U.S. News & World Report*, August 2, 46–55.

Wysocki, Jr., B.J. (2004). Industrial strength: To fix health care, hospitals take tips from factory floor. *The Wall Street Journal*. April 9, A1, A9.

KEY TERMS

Database: One possible definition is that a database is a collection of records stored in a computer in a systematic way, so that a computer program can consult it to answer questions. For better retrieval and sorting, each record is usually organized as a set of data elements (facts). The items retrieved in answer to queries become information that can be used to make decisions. The computer program used to manage and query a database is known as a database management system (DBMS). The properties and design of database systems are included in the study of information science.

Digital: The word *digital* is most commonly used in computing and electronics, especially where real-world information is converted to binary numeric form, as in digital audio and digital photography. Such data-carrying signals carry either one of two electronic or optical pulses, logic 1 (pulse present) or 0 (pulse absent). The term is often used with the prefix "e-" as in e-mail and e-book, even though not all electronics systems are digital.

Information Systems: The study of information systems is usually a commerce and business administration discipline, and frequently involves software engineering, but it also distinguishes itself by concentrating on the integration of computer systems with the aims of the organization. The area of study should not be confused with computer science, which is more theoretical in nature and deals mainly with software creation, or computer engineering, which focuses more on the design of computer hardware.

Information Technology: IT, also known as Information and Communication(s) Technology (ICT) and Infocomm (especially in Asia) is a broad subject concerned with technology and other aspects of managing and processing information, especially in large organizations. In particular, IT deals with the use of electronic computers and computer software to convert, store, protect, process, transmit, and retrieve information. For that reason, computer professionals are often called IT specialists, and the division of a company or university that deals with software technology is often called the IT department. Other names for the latter are information services (IS), management information services (MIS), or managed service providers (MSP).

Network Card: A network card, network adapter, or network interface card (NIC) is a piece of computer hardware designed to allow computers to communicate over a computer network.

Telemedicine: Telemedicine is the delivery of medicine at a distance. The term is composed of the Greek word *τελε* (*tele*) meaning "far," and *medicine*. Telemedicine may be as simple as two health professionals discussing a case over the telephone, or as complex as using satellite technology and videoconferencing equipment to conduct a real-time consultation between medical specialists in two different countries. Telemedicine generally refers to the use of communications and information technologies for the delivery of clinical care.

Development of Walking Pattern Evaluation System for Hypogravity Simulation

R. Leães
Pontifical Catholic University, Brazil

R. Cambraia
Pontifical Catholic University, Brazil

F. Bacim
Pontifical Catholic University, Brazil

G. Dalmarco
Pontifical Catholic University, Brazil

A. Calder
University of Glasgow, Scotland

D. F. G. De Azevedo
Pontifical Catholic University, Brazil

M. Pinho
Pontifical Catholic University, Brazil

T. Russomano
Pontifical Catholic University, Brazil

INTRODUCTION

The acceleration due to gravity at the surface of a planet varies directly as the mass and inversely as the square of the radius. This follows directly from Newton's law of universal gravitation.

The Moon is 384,403km distant from the Earth. Its diameter is 3,476km. The acceleration due to gravity is 1.62 m/s^2 because the Moon has less mass than Earth. It is approximately 1/6 that of the acceleration due to gravity on Earth, 9.81 m/s^2.

Mars and Earth have diameters of 6,775km and 12,775km, respectively. The mass of Mars is 0.107 times that of Earth. This makes the gravitational acceleration on Mars $3.73 m/s^2$, as expressed on Equation 1:

$$gm = 9.8 \times 0.107 \times (12775/6775)^2 = 3.73 m/s^2 \qquad (1)$$

Therefore, if a body weighs 200N on Earth, it is possible to calculate how much it would weigh on Mars. Knowing that the weight of an object is its mass (m) times the acceleration of gravity, we can have W = m x g; 200 = 9.8 x m; m = 20.41kg. This mass is the same on Mars, so the weight on Mars is Wmars = 3.73 x 20.41 = 76.1N.

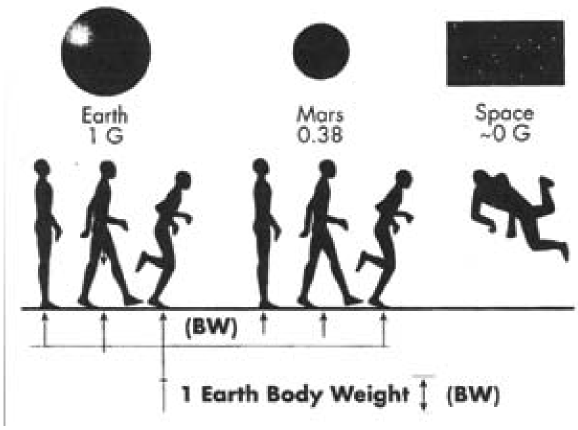
Figure 1 illustrates the change in body posture of an individual on Earth, on Mars, and in space (microgravity).

BACKGROUND

There are three primary techniques for simulating partial gravity: water immersion (neutral buoyancy), parabolic flight, and **body suspension device** (BSD) models.

Underwater Immersion. During tests, a neutrally buoyant subject is ballasted to simulate the desired partial gravity loading. For example, one-sixth of the subject's body mass is added in ballast if a lunar simulation is desired. Water immersion offers the subject freedom from time constraints and freedom of movement, but the hydrodynamic drag is disadvantageous for movement studies.

Figure 1. Schematic representation of body posture in different gravitational environments. (www.nasa.gov)



Parabolic Flight. NASA KC-135 aircraft or Russian IL-76 aircraft are typically used to simulate partial gravity by flying Keplerian trajectories through the sky. This technique provides approximately 30s and 40s for lunar gravity and Martian gravity tests, respectively. Parabolic flight is the only way to effect true partial gravity on Earth, but experiments are expensive and of limited duration.

Body Suspension Device (BSD) Models. BSDs have been developed and used worldwide to study human physiology during partial or total simulation, such as the Zero Gravity Locomotion Simulator (ZLS) (Cavanagh, McCrory, Baron & Balkin, 2002) and the Zero Gravity Simulator (ZGS) (D'Andrea & Perusek, 2005).

The Cleveland Clinic Foundation and Penn State University built the ZLS, which is mounted vertically in a freestanding frame and includes padded straps that support a runner under the head, torso, arms, and legs. In this position, there is no gravitational force between the runner and the machine.

In physiotherapy, body suspension devices provide patient and therapist security, giving increased freedom of movement to accomplish technical and facilitative maneuvers. Manual assistance is necessary to promote the postural adjustment in patients with neurological injury that uses a gait training system (Boehrman & Harkema, 2000).

Virtual reality offers potential for the development of evaluation and training systems, allowing precise control of a stimulus. Researchers have utilized gait training on a treadmill body suspension system with virtual reality. They found that people with multiple sclerosis increased the gait velocity, and their balance was improved after two months of treatment with the Body Weight Suspension and Treadmill (BWS/TM) system (Fulk, 2005). The virtual reality technology employed was based on navigation. Three virtual environments representing the soils of Earth, Mars, and the Moon were developed. The hypotheses were (1) the reduction of the gravitational force alters gait kinematic parameters; (2) the utilization of a Head Mounted Display (HMD), virtual reality glasses, while walking on a treadmill influences the postural balance of the user.

MAIN FOCUS OF THIS STUDY

The main objective of this research was to develop a secure and efficient system for **gait evaluation** that can be used for research purposes in aerospace medicine, physiotherapy, computer science, and rehabilitation medicine. A system was developed to evaluate **walking patterns** in **hypogravity simulation**, which was called SAMSH. This consisted of the improvement of a body suspension device and the instrumentation of a treadmill that served as a platform to assess a physical locomotion technique in a virtual environment.

Details of the equipment developed for this study are presented as follows:

The body suspension device (3000 mm x 2660 mm wide and 2000mm high) has a counterweight system of 20 bars (5kg each). The subject wore a harness (Advanced Air Sports Products; Lake Elsinore, CA), which was suspended by a steel wire and designed to support massive tensions (Figure 2). A load cell permitted the measurement of the mechanical stress by means of Wheatstone bridge.

Equation 2 demonstrates how to calculate the relative mass of a subject in a simulated gravitational field, where RM = relative mass (kg), BM = body mass on Earth (kg), SGF = simulated gravitational force (m/s^2) and 1G = 9.81 m/s^2.

Equation 3 gives the counterweight (CW, kg) necessary to simulate body mass for a preset hypogravity level.

$$RM = \frac{BM \times SGF}{1G} \qquad (2)$$

$$CW = BM - RM \quad (3)$$

A moving platform was based on the instrumentation of an electrical treadmill. The velocity of the treadmill was controlled using a communication protocol with a frequency inverter Movtrack07 (SEW). It was also necessary to develop an electrical interface RS232/485 to provide communication between the computer and the inverter. Treadmill velocity was at 40m/min throughout the tests (Figure 3).

The toolkit Small VR (developed by the Virtual Reality Group of PUCRS, based on OpenGL and GLUT libraries) was developed to represent the Martian, Terrestrial, and Lunar surfaces. MS visual C++ 6.0 software (C and C++ Programming Language) was used. The route followed by the subject was predetermined. The subject was instructed to walk forward at 40m/min for approximately 2.15min. Dimensions for the virtual environment vertices were set at 10m by 10m. A proportional equation allowed the individual's height to be kept the same independent of his position, creating a realistic environment. Soil textures of Earth, Mars, and the Moon were extracted from graphic computation libraries (Figure 4).

a

Figure 2. Body suspension device: counterweight system (a); harness used (b); the load cell (c)

Figure 3. Schematic view of the electronic velocity control of the treadmill

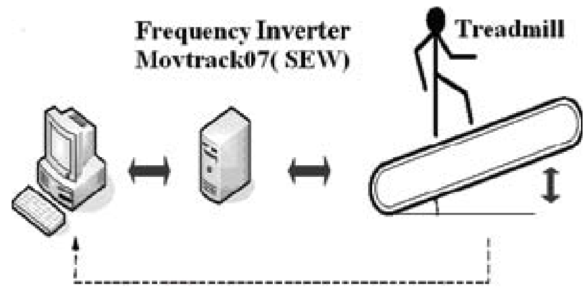
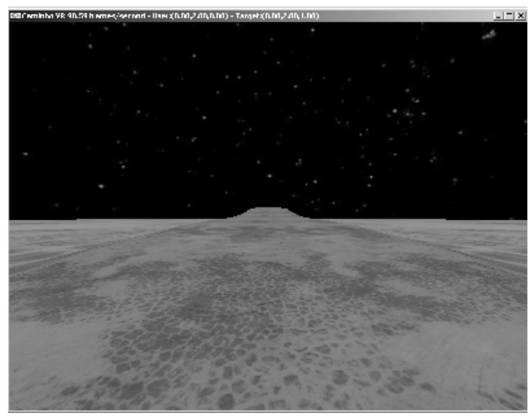


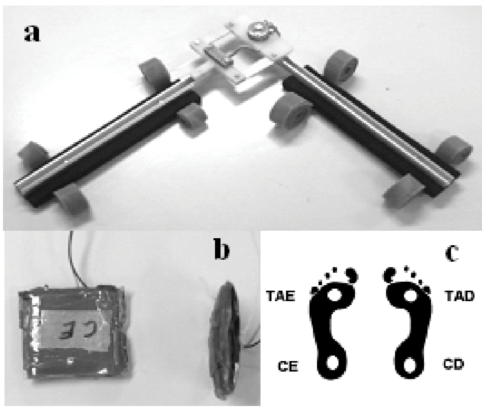
Figure 4. Virtual Mars environment



Kinematic gait was performed using spatial and temporal patterns. Electrogoniometers were used to evaluate the angular displacement of the knee and ankle (Figure 5a). Each electrogoniometer consists of two aluminium bars covered with rubber material and connected by a 10KΩ potentiometer. They were fed by a 5V power source. Calibration was performed before the beginning of each experiment using a manual goniometer. Data was acquired by means of DataQ. The linearity of the electrogoniometers has been evaluated previously (Gurgel, Porto, Russomano, Castro, Bertoglio & Schroeder, 2004).

Foot switches were developed for the measurement of temporal parameters that would determine the phases of the locomotion cycle. When a region of the foot was in contact with the surface, a short circuit supplied a level of tension that was captured by the DataQ and stored for subsequent processing (Figures 5b and 5c).

Figure 5. Electrogoniometer (a); Foot switches placed on the front (TAE, TAD) and back (CE, CD) part of the plantar region (b and c)



The study protocol was approved by the PUCRS Research Ethics and Scientific Committees. Each subject provided written informed consent before participating in the experiment.

A preliminary assessment of the system was conducted with the participation of one subject. The subject experienced two different conditions of body weight reduction: 30% for Mars and 60% for the Moon. Descriptive and angular variables related to kinematic gait evaluation were measured. Displacement of the subject's head into the three planes of movement was measured using the head mounted display (HMD) on x, y, and z axes (Pitch, Yaw, and Roll). The length, duration, speed of the stride, cadence, and duration of the stance phase were recorded.

VARIABLES EVALUATED DURING WALKING ON MARTIAN AND LUNAR GROUND-BASED SIMULATION BY MEANS OF A BODY SUSPENSION MODEL

The stride time (TS) is the time between the first and the second contact of the same foot region on the ground. The stride length can be calculated using the equation ($v = d/t$). Cadence was determined by the number of steps per minute of the gait cycle. The stance phase could be determined by knowing the contact time (CT) of the feet regions instrumented by the foot switches. It was observed from the time between the initial and the final contact of each region (front and back part of the plantar region) of each foot during a stride. The aerial time (AT) is the time between the toe-off and ground contact of the opposite foot, representing the swing phase. This is more significant in hypogravity environments than at 1G (Newman, 2000).

Angular variables were evaluated by the electrogoniometers, which determine the ankle dorsal flexion (0º to 30º) and ankle plantar flexion (0º to 50º), flexion (-10º to 90º) and extension (90° to -10º) of the knee.

The statistical analysis was made using SPSS (version 10.0), ANOVA, and t-tests. Significance level adopted was about 5%.

The findings of this study demonstrated that SAMSH adequately simulates the gravitational forces of Mars and the Moon for the study of the walking patterns in hypogravity environment, and the use of electrogoniometers and foot switches allowed the evaluation of walking variables.

Stride time, length stride, and stance phase increased progressively from 1G to Mars simulation and then the Moon simulation. The opposite was observed in the cadence, which was higher at 1G and lower on the Moon simulation. Aerial time was 0,69s on the Moon, 0,63s on Mars, and 0,50s on Earth.

Ankle dorsal flexion at Earth was 29º, increasing to 69.7º on Mars and to 67.9º on the Moon simulations. The ankle plantar flexion showed an opposite pattern, being bigger on the Moon and lower on the Earth. Flexion of the knee was 93.6º on the Moon simulation and 66.8º at Earth. The extension of knee was bigger on Earth (61.8º) and lower on the Moon (28.4º).

HMD data suggested that head movement was more prominent at the X axis (pitch, head up and down) on the Earth, Y axis (yaw, right and left inclination) on Mars simulation, and Z axis (roll, left and right rotation) on the Moon simulation.

FUTURE TRENDS

The era of manned space flight began on April 12, 1961, with the launch of Yuri Gagarin aboard Vostok-1 by the former Soviet Union for a very short space mission that lasted less than 2h. This historic mission was followed by a series of American- and Soviet-manned space programs. Space tourism, a recent phenomenon, is the term that has come to be used to mean ordinary members of the public buying tickets to travel to space and back. They have been called space tourists, citizen explorers, public spaceflight participants, and astourists.

Tourism in space has been considered a fledgling industry, born out of necessity, yet driven by the same curiosity and ambition that took humanity to the Moon. In Russia, Europe, and the United States, private companies are already vying to become space tourism leaders.

It was clear since the first manned space flight that the main objectives of space life science research are to adapt man to the environment of space and readapt him when back to Earth. The evaluation of human walking patterns in hypogravity environments is then mandatory.

CONCLUSION

According to the present study, the greater the reduction of body weight, the longer the stride time. This reflects findings in previous other studies (Newman, 2000). Further, the time of contact diminishes with body weight reduction.

During the simulation of the Moon, steps were longer and less frequent than at 1G. A similar change in locomotion has been noted in aquatic environments (Degani & Barela, 2001). Reduction in subject body weight also influences ground reaction force (Amadio, 1996). There is a significant reduction in the peaks of force during locomotion in partial gravity (Newman, 2000).

Qualitative and quantitative data obtained by HMD and video cameras showed that head movements were predominantly on the z axis. Lateral trunk inclination was observed during body weight reduction (30% and 60%) as a result of limit movements of the trunk and pelvic region. This reflects findings of Rose and Gamble (1998), who observed increased lateral trunk inclination in proportion with increases in stride length. A protective extension reaction, a postural reflex to avoid lost of body balance, was also observed with the utilization of HMD (Bobath, 1978).

Kinematic evaluation of gait variables on Earth without HDM (control) were speed 0.67 m/s, cadence 86 steps/min, length of stride 96 cm, knee angular variation 66.8º (flexion) and 5.2º (extension), ankle angular variation 27º between dorsal flexion (73º) and plantar flexion (100º). These results are similar to the ones obtained by Winter (1991), who states that the normal speed at 1G should range from 1.16 m/s to 1.67 m/s. However, the gait speed in this research was slower in order to allow the subject to use the HDM safely. No data regarding walking patterns with virtual reality were found in the literature for comparison.

REFERENCES

Amadio, A.C. (1996). *Fundamentos biomecânicos para a análise do movimento humano*. São Paulo: Laboratório De Biomecânica /Eefusp.

Bobath, B. (1978). Atividade reflexa postural anormal, 2ª Edição. São Paulo: Editora Manole.

Boehrman, A.L., & Harkema, S.J. (2000). Locomotor training after human spinal cord injury: A series of case studies. *Phys Ther, 80,* 688–700.

Cavanagh, P.R., McCrory, J.L., Baron, H.A., & Balkin, S. (2002). Locomotion in simulated microgravity: Gravity replacement loads. Avia*tion, Space, and Environmental Medicine, 73*(7).

D'Andrea, S., & Perusek, G. (2005). Jumping in a simulated and true microgravity: Response to maximal efforts with three landing types. *Aviation, Space, and Environmental Medicine, 76*(5).

Degani, A.M., & Barela, J.A. (2001). Parâmetros espaço-temporais e angulares do andar de indivíduos idosos em ambiente aquático. In Congresso Brasileiro de Biomecânica, 9. Gramado. Anais. Gramado: Ufrj, 152–157.

Fulk, G.D. (2005). Locomotor training and virtual reality-based balance training for an individual with multiple sclerosis: A case report. *Journal of Neurologic Physical Therapy*.

Gurgel, J., Porto, F., Russomano, T., Castro, L., Bertoglio, R., & Schroeder, I. (2004). *Construção e calibração de eletrogoniômetro de baixo custo, baseado em potenciômetro, para analise biomecânica do movimento humano*. Proceedings of the Anais Xxvii, Simpósio Internacional De Ciência Espacial, São Paulo, 256.

Newman, D.J. (2000). Life in extreme environments: How will humans perform on Mars? *Gravitational and Space Biology Bulletin, 13*(2).

Rose, J., & Gamble, J. (1998). *Marcha humana*. São Paulo: Editorial Premiere.

Winter, D.A. (1991). *The biomechanics and motor control of human gait: Normal, elderly and pathological* (Second Ed.). University of Waterloo Press.

Diagnostics, Therapeutics, and Health Informatics in Osteoporosis

Sarah H. Gueldner
Decker School of Nursing, USA

Guruprasad Madhavan
Clinical Science and Engineering Research Center, USA

Eric D. Newman
Geisinger Health System, USA

Carolyn S. Pierce
Decker School of Nursing, USA

INTRODUCTION

Osteoporosis, usually silent until a fracture occurs, is among the most common health problems facing elders worldwide. By definition, osteoporosis is a "systemic" skeletal disease characterized by a low bone mass and a micro architectural deterioration of bone tissue, with a consequent increase in bone fragility and susceptibility to fracture ("*The Burden of Musculoskeletal Conditions*," 2003). The incidence of osteoporotic hip fracture increases exponentially with age, and the increase in older persons globally could dramatically increase the number of hip fractures, posing a devastating increase in disability and cost for elders worldwide. Therefore, it is imperative that diagnostic and treatment measures be developed and instituted worldwide to support preventative measures for osteoporosis and consequential fractures. Toward that purpose, the World Health Organization (WHO) has declared 2002–2011 as the *Decade of the Bone and Joint*, uniting nations throughout the world in the commitment of energy and resources to accelerate progress in bone health and prevention of fractures. Keeping in mind this global context, this discussion includes information about the prevalence and impact of osteoporosis, its signature pathology (including bone remodeling), factors which place individuals at risk for developing osteoporosis, and the role of diagnostics, therapeutics, and informatics in the realm of osteoporosis. Encouraging information is also provided about recent innovative technological developments that may enhance our ability to detect and treat osteoporosis earlier, in time to reduce and better manage its unwelcome sequelae.

PREVALENCE AND IMPACT

The WHO estimates that up to 70% of women older than 80 years of age have osteoporosis ("*The Burden of Musculoskeletal Conditions*," 2003). One in every two women and one in every five men over 50 years of age experience an osteoporotic fracture during their lives, and more women die from the after-effects of osteoporotic fractures than from cancer of the ovaries, cervix, and uterus together. There are 1.5 million osteoporotic fractures in the U.S. each year, 20% hip fractures, 50% vertebral fractures that lead to height loss, and 30% at the wrist and other sites. One fourth (25%) of all postmenopausal women will develop a vertebral deformity and another 15% will sustain a fractured hip, numbers that could triple by 2040 ("*Bone Health and Osteoporosis: A Report*," 2004).

Hip fractures are the most severe consequence of osteoporosis, and are associated with lengthy hospital admissions, difficulty in performing activities of daily life, nursing home placement, and a high rate of mortality. The annual worldwide incidence of hip fracture is projected to grow to 2.6 million by 2025 and to 4.5 million by 2050. Virtually all persons in the U.S. with hip fractures are hospitalized, and 25% of those who sustain hip fractures require long-term nursing home care. Spine fractures are considerably less problematic in terms of cost, with only 10% requiring hospitalization and less than 2% being admitted to a nursing home. However, they account for 66,000 physician office visits and at least 45,000 hospital admissions each year.

The economic burden of osteoporosis to society is immense. Each year in the U.S., osteoporotic fractures result in more than 500,000 hospitalizations, 800,000 emergency room visits, more than 2.6 million physician office visits, and the placement of nearly 180,000 individuals in nursing homes. It is estimated that each hip fracture represents approximately $40,000 in total medical costs. Care to white women comprises the majority of these costs, but men and nonwhite women of all ages account for $1 billion in direct costs of osteoporotic fractures. Due to the late age at which the expenses related to care occur, osteoporosis accounts for 14% of all nursing home days. Given that 75% of all hip, spine, and distal forearm fractures occur in persons 65 years and older, the direct costs are largely borne by society, in the form of social reimbursement programs. In the United States, Medicare pays approximately one half of these costs, and Medicaid covers an additional one-fourth of the expense.

But the impact of osteoporosis on the personal lives of the patients and their families is even greater. Nearly one in five persons who sustain a hip fracture end up in a nursing home, and about 20% will die before a year has passed. Two-thirds of hip fracture patients never return to their prefracture level of function, and many lose their ability to walk. More than one-third (37%) lose their ability to dress themselves, 58% lose their ability to move from their chair to standing, and 60% lose their capacity to walk across the room. Approximately half of the individuals who sustain hip fractures never walk independently again, even if they were ambulatory before their fracture (Gueldner, Burke, Wright, & Newman, 2007).

PATHOLOGY

Osteoporosis is not only the result of accelerated bone loss during aging, but may also develop because of suboptimal bone growth in childhood and adolescence, or by the use of certain medications (most notably steroids) for the treatment of other health problems. Fracture risk is currently estimated by measuring bone mineral density (BMD), with the risk of fracture increasing sharply as BMD declines. The estimated life-time risk for wrist, hip, and vertebral fractures is 15%. Hip fracture rate is used to calculate the osteoporosis fracture burden. Prior fracture is one of the most important predictors of future fracture. The risk of hip fracture after a wrist fracture increases 1.4-fold in U.S. women, 1.5-fold in Swedish women, and 1.8-fold in Danish women. Wrist fracture is an even stronger predictor of hip fracture in men; U.S. men who had a wrist fracture were found to be 2.3 times more likely to sustain a hip fracture, and Swedish men with wrist fracture were 2.8 times more likely to sustain hip fracture.

Although an individual's height and unique bone structure are developed by late adolescence, new bone continuously replaces old bone throughout life in a process referred to as bone remodeling. Through the remodeling process, approximately 10% of the skeleton is replaced each year, with most of the adult skeleton being replaced every 10 years. The skeleton consists of 80% cortical (compact) bone, which is located predominately on the outer surfaces of the bone, and the remaining 20% is inner trabecular (spongy) tissue, which forms a large portion of the vertebrae, proximal femur, and distal radius. If less new bone is formed during the remodeling process than the amount of old bone removed, a net loss of bone occurs, and the resulting fragile bone is more likely to break with less trauma, as is the clinical problem with osteoporosis. Remodeling occurs in both types of bone, but most (80%) occurs in the trabecular bone. Thus fractures in regions of high trabecular area such as the hip, spine, and wrist are more likely to occur.

In the complex bone remodeling sequence, bone cells lie quiescent until osteoclasts (bone resorbing cells) are stimulated to resorb (remove) a small volume of bone. While some details of the actual remodeling process remain unclear, it is thought that this resorbing activity is increased by a number of factors, including parathyroid hormone (PTH) and thyroxine, and that it is decreased by estrogen, testosterone, vitamin D, calcium, high phosphorus levels, and sometimes by cytokines from osteoblasts (bone forming cells). Localized prostaglandins and interleukins may also exert a stimulatory or inhibitory effect on this remodeling process. A low calcium intake stimulates the secretion of PTH, which in turn activates the osteoclasts to release lysosomal enzymes which digest bone matrix, causing the release of calcium and other bone minerals and proteins. After the resorption cavity is formed, a cement line is laid down by macrophage-type mononuclear cells, limiting further resorption of bone in that particular area. The cement line is rich in osteopontin, which may serve to shut off osteoclast (resorption) activity and stimulate osteoblast (rebuilding) activity. A growing number of pharmaceutical modalities have been developed with the goal of reducing bone loss or increasing bone formation to maintain bone health by inhibiting osteoclast action on bone resorption, or by allowing formation to continue.

DIAGNOSTIC TECHNOLOGIES IN OSTEOPOROSIS

Osteoporotic fractures are preventable, but appropriate diagnostic tools are crucial in the early assessment and treatment of osteoporosis. Bone densitometry has made it possible for health care providers to diagnose osteoporosis before the first fracture occurs; and to reliably predict fracture risk in postmenopausal women and others at high risk. It can also be used as a surrogate marker to determine the effectiveness of therapies and to evaluate those patients that might not respond to osteoporosis therapy. BMD measurements correlate strongly with the load bearing capacity of the hip and spine, and can detect osteoporosis before a fracture occurs, predict fracture risk, assess the rate of bone loss with repeated measures, and monitor the effectiveness of treatment for osteoporosis. The two technologies most frequently used for bone density testing are heel ultrasound (HUS) and dual x-ray absorptiometry (DXA).

HUS uses change in speed of sound and waveform attenuation to create a number that correlates with bone density. A T-score is created by comparing this extrapolated value to a normative database, and this value has been shown to predict fracture risk at the hip and spine. The WHO classifies normal bone density as a T-score of -1 or above, osteopenia as a T-score between -1 and -2.5, and osteoporosis T-score -2.5 or below ("*The Burden of Musculoskeletal Conditions,*" 2003). HUS has some chief advantages, including portability, no radiation, and low cost. This makes it an appealing technology for patients as well as primary care offices. But a major disadvantage to HUS is that a significant proportion of patients will need a DXA in addition, for two reasons, including indeterminate results and a need for ongoing monitoring of low bone density or treatment effectiveness.

On the other hand, DXA is currently considered the "gold-standard" fracture risk assessment tool. Used to measure BMD in the hip, spine, or wrist, the DXA takes about 10 minutes to perform with low radiation exposure (one-tenth of a standard chest x-ray). The results are also reported in T-scores. However, DXA is not universally available for screening for osteoporosis, partly because the equipment is large and expensive, and is not transportable, which often poses a barrier to access for persons who live in nonurban areas. However, mobile DXA modalities have just started to emerge and their large-scale adoption and long-term feasibility remain to be seen (Newman, Olenginski, Perruquet, Hummel, Indeck, & Wood, 2004).

THERAPEUTIC TECHNOLOGIES IN OSTEOPOROSIS

In terms of preventative treatment technologies, padded hip protectors had received some earlier attention based on their promise for decreasing fracture during falls in elderly populations, but recent evidences seem to point that their value in fracture prevention in persons with severe osteoporosis is minimal (Parker, Gillespie, & Gillespie, 2006). There are also a growing number of nonpharmaceutical muscle calf muscle pump stimulation modalities under study and development that hold potential for helping to reduce bone loss and stimulate bone growth exogenously. One such modality is direct electrical stimulation, whose efficacy in preventing bone loss and even increase bone mass, by means of calf muscle pump activation, has been widely studied in individuals with spinal cord injury (Belanger, Stein, Wheeler, Gordon, & Leduc, 2000; Claydon, Steeves, & Krassioukov, 2006; Eser, de Bruin, Telley, Lechner, Knecht, & Stussi, 2003; Sampson, Burnham, & Andrews, 2000). Current clinical evidence suggests that, at least conceptually, direct electrical stimulation of the musculature may have potential as a means to prevent bone loss. However, electrical stimulation is not conveniently applied, can be painful or induce discomfort, and often leads to rapid muscle fatigue, factors that may significantly limit its applicability as a long-term prevention strategy for osteoporosis.

As a means of bypassing the complications of direct electrical stimulation of muscle contraction, investigators have recently begun pursuing the concept of using reflex-mediated pathways to indirectly trigger muscle activity. Micromechanical stimulation of the plantar surface stimulates the cutaneous somatosensory mechanoreceptors, which subsequently initiate calf muscle contraction (Kennedy & Inglis, 2002; Madhavan, Stewart, & McLeod, 2004). This strategy has been implemented in a device which can be placed in front of a chair, or under a desk, so that the user can readily obtain calf muscle pump stimulation in a sustained manner, in either the home or workplace.

Studies utilizing this technology have shown that calf muscle pump activation can be readily enhanced, with significant increases in lower body fluid flow (Stewart, Karman, Montgomery, & McLeod, 2005) and venous return from the lower limbs, resulting in reversal of orthostatic hypotension and orthostatic tachycardia (Madhavan, Stewart, & McLeod, 2005). Consistent with these observed effects on lower limb muscle pump activity, sustained plantar stimulation has been shown to significantly increase lower limb muscle strength in humans (Russo et al., 2003; Torvinen et al., 2003). Correspondingly, these effects on the musculature have been observed to affect bone density over the long term in animal models (Rubin, Turner, Bain, Mallinckrodt, & McLeod, 2001), children (Ward, Alsop, Caulton, Rubin, Adams, & Mughal, 2004), and adults (Rubin, Pope, Fritton, Magnusson, Hansson, & McLeod, 2003; Rubin, Recker, Cullen, Ryaby, McCabe, & McLeod, 2004). These preliminary results, combined with the ease of use of this technology, suggest that this technology may, in the near future, form the basis of a convenient, noninvasive, nonpharmacologic means to prevent or reduce age-related bone loss and osteoporosis.

HEALTH INFORMATICS IN OSTEOPOROSIS

The United States health care infrastructure is increasingly recognizing the potency of information technology in significant cost savings (Hillestad et al., 2005; Walker, 2005), and health outcomes ("*Crossing the Quality Chasm*," 2001), especially for the care of the elderly population. Health information technology has relevant application to osteoporosis, both in terms of providing educational information to increase the public's awareness and knowledge of lifestyle habits that promote bone health, and in providing system linked quick reference information and pop up medical records and reminders for primary health care providers as they see their patients.

The Institute of Medicine has declared information technology as the structural component of health care most likely to improve the process of health care delivery ("*The Computer-Based Patient Record*," 1997). There has long been consensus that accurate and consistent use of a computer by the clinician is essential to the monitoring of outcomes, and the recommended standards were quickly exceeded within general practice. However, the literature confirms that the quality of the data seen in such innovative systems has not yet occurred in relation to the prevention of osteoporotic fractures, which is surprising, given the extraordinary health care costs associated with fragility fractures (Bayly, 2005). Rather, incident fractures are not being commonly managed systematically in either secondary or primary care (Chen, Brown, Archibald, Aliotta, & Fox, 2000).

It has been pointed out that general practice is probably best placed to manage those at risk, because it is they who see the patients in time to institute early diagnosis and preventative measures. However, data quality in primary care regarding osteoporosis has been found to be lacking and inconsistent, particularly in regard to data related to fracture. Thus front line clinicians represent a fertile area for instituting improvements in both availability and use of medical informatics. This can be accomplished through two mechanisms: the acquisition of clinical information and the process of monitoring the continued delivery of care by electronic rather than paper-based ways.

TECHNOLOGY NEEDS OF THE FUTURE

Worldwide, the number of persons who sustain hip fractures every year is more than 1.3 million people, and that number is predicted to double by 2020. This dramatic increase in hip fractures alone will have a devastating effect on disability and cost for elders worldwide. However, other important chronic diseases are also projected to increase due to the significantly aging society (including dementia, cancer, cardiovascular disease, and diabetes), and some of these conditions already generate more cost and disability than musculoskeletal disorders, chiefly osteoporosis. And in some areas of the world, acute health problems are so pressing that little attention can be paid to insidious conditions such as osteoporosis. For instance, the 70 million people estimated by WHO to have osteoporosis pales when compared with the two billion persons infected with tuberculosis (McGovern & Bond, 2003), three million deaths each year from diarrheal diseases (Murray & Lopez, 1996), and 1 million deaths from malaria (Gelb & Hol, 2002). Given the simultaneous worldwide toll from these widespread and equally devastating diseases, it seems that osteoporosis preventative strategies will have to rely primarily on cost-effective public health measures. And it is in this aspect that the fields of engineering and information technology can and must take the lead, through the development of cost-effective diagnostic, therapeutic, and information system solutions. Representative areas of applications include provision of patient-tailored reminders related to the skeletal health status, with indications to diagnostic measures, customized therapeutic options, and portable electronic preventative health care records. Further, it is equally vital that sophisticated but user friendly health informatics be developed to provide educational outreach about bone health to individuals, families, and the community.

CONCLUSION

Osteoporosis is a devastating public health problem that affects all strata of the global community. But healthy women and men in the age range of 50-65 years still have time to engage in osteoporosis preventing behaviors to minimize bone loss and fractures. It is critical that concerted efforts be directed toward community education and the development of new diagnostics, therapeutics, and informatics to liberate towns and communities worldwide of the ravages of osteoporosis, as our predecessors have spared modern society from the devastation of polio and rickets. Toward this goal, it is imperative that additional research be conducted to find cost-effective diagnostic, treatment, and information technology measures, in order that both professionals and the general public can be better informed about preventative aspects of musculoskeletal health care.

REFERENCES

Bayly, J. (2005). IT and osteoporosis - the GP's perspective. *Osteoporosis Review, 13*(3), 15-18.

Belanger, M., Stein, R. B., Wheeler, G. D., Gordon, T., & Leduc, B. (2000). Electrical stimulation: Can it increase muscle strength and reverse osteopenia in spinal cord injured individuals? *Arch Phys Med Rehabil, 81*(8), 1090-1098.

Bone health and osteoporosis: A report of the surgeon general. (2004): Rockville, MD: US Department of Health and Human Services, Office of the Surgeon General.

The burden of musculoskeletal conditions at the start of the new millennium: Report of the world health organization scientific group (2003). Geneva, Switzerland.

Chen, A., Brown, R., Archibald, N., Aliotta, S., & Fox, P. (2000). *Best practices in coordinated care*. Princeton, NJ: Mathematica Policy Research, Inc.

Claydon, V. E., Steeves, J. D., & Krassioukov, A. (2006). Orthostatic hypotension following spinal cord injury: Understanding clinical pathophysiology. *Spinal Cord, 44*(6), 341-351.

The computer-based patient record: An essential technology for health care (Rev. Ed.). (1997). Washington, DC: Institute of Medicine, National Academy Press.

Crossing the quality chasm: A new health system for the 21st century. (2001). Washington, DC: National Academy Press.

Eser, P., de Bruin, E. D., Telley, I., Lechner, H. E., Knecht, H., & Stussi, E. (2003). Effect of electrical stimulation-induced cycling on bone mineral density in spinal cord-injured patients. *Eur J Clin Invest, 33*(5), 412-419.

Gelb, M. H., & Hol, W. G. (2002). Parasitology. Drugs to combat tropical protozoan parasites. *Science, 297*(5580), 343-344.

Gueldner, S., Burke, M., Wright, H., & Newman, E. (2007). *Preventing and managing Osteoporosis*. New York: Springer Publishing Company.

Hillestad, R., Bigelow, J., Bower, A., Girosi, F., Meili, R., Scoville, R., et al. (2005). Can electronic medical record systems transform health care? Potential health benefits, savings, and costs. *Health Aff (Millwood), 24*(5), 1103-1117.

Kennedy, P. M., & Inglis, J. T. (2002). Distribution and behaviour of glabrous cutaneous receptors in the human foot sole. *J Physiol, 538*(Pt. 3), 995-1002.

Madhavan, G., Stewart, J. M., & McLeod, K. J. (2004). Mechanosensory characterization of plantar vibration effects on the skeletal muscle pump activity (Abstract). *The FASEB Journal, 18*(4), A648.

Madhavan, G., Stewart, J., & McLeod, K. (2005). Effect of plantar micromechanical stimulation on cardiovascular responses to immobility. *Am J Phys Med Rehabil, 84*(5), 338-345.

McGovern, V., & Bond, Q. (2003). Global health research. *Science, 300*(5628), 2003.

Murray, C., & Lopez, A. (Eds.). (1996). *The global burden of disease and injury series, Volume II. Global health statistics: A compendium of incidence, prevalence and mortality estimates for over 200 conditions.* Cambridge, MA: Harvard University Press.

Newman, E. D., Olenginski, T. P., Perruquet, J. L., Hummel, J., Indeck, C., & Wood, G. C. (2004). Using mobile DXA to improve access to osteoporosis care: Unit design, program development, implementation, and outcomes. *J Clin Densitom, 7*(1), 71-76.

Parker, M. J., Gillespie, W. J., & Gillespie, L. D. (2006). Effectiveness of hip protectors for preventing hip fractures in elderly people: Systematic review. *Bmj, 332*(7541), 571-574.

Rubin, C., Pope, M., Fritton, J. C., Magnusson, M., Hansson, T., & McLeod, K. (2003). Transmissibility of 15-hertz to 35-hertz vibrations to the human hip and lumbar spine: determining the physiologic feasibility of delivering low-level anabolic mechanical stimuli to skeletal regions at greatest risk of fracture because of osteoporosis. *Spine, 28*(23), 2621-2627.

Rubin, C., Recker, R., Cullen, D., Ryaby, J., McCabe, J., & McLeod, K. (2004). Prevention of postmenopausal bone loss by a low-magnitude, high-frequency mechanical stimuli: A clinical trial assessing compliance, efficacy, and safety. *J Bone Miner Res, 19*(3), 343-351.

Rubin, C., Turner, A. S., Bain, S., Mallinckrodt, C., & McLeod, K. (2001). Anabolism. Low mechanical signals strengthen long bones. *Nature, 412*(6847), 603-604.

Russo, C. R., Lauretani, F., Bandinelli, S., Bartali, B., Cavazzini, C., Guralnik, J. M., et al. (2003). High-frequency vibration training increases muscle power in postmenopausal women. *Arch Phys Med Rehabil, 84*(12), 1854-1857.

Sampson, E. E., Burnham, R. S., & Andrews, B. J. (2000). Functional electrical stimulation effect on orthostatic hypotension after spinal cord injury. *Arch Phys Med Rehabil, 81*(2), 139-143.

Stewart, J. M., Karman, C., Montgomery, L. D., & McLeod, K. J. (2005). Plantar vibration improves leg fluid flow in perimenopausal women. *Am J Physiol Regul Integr Comp Physiol, 288*(3), R623-629.

Torvinen, S., Kannus, P., Sievanen, H., Jarvinen, T. A., Pasanen, M., Kontulainen, S., et al. (2003). Effect of 8-month vertical whole body vibration on bone, muscle performance, and body balance: A randomized controlled study. *J Bone Miner Res, 18*(5), 876-884.

Walker, J. M. (2005). Electronic medical records and health care transformation. *Health Aff (Millwood), 24*(5), 1118-1120.

Ward, K., Alsop, C., Caulton, J., Rubin, C., Adams, J., & Mughal, Z. (2004). Low magnitude mechanical loading is osteogenic in children with disabling conditions. *J Bone Miner Res, 19*(3), 360-369.

KEY TERMS

DXA: Dual x-ray absorptiometry.

Hip Fracture: Broken hip.

HUS: Heel ultrasound.

Inner Trabecular Tissue: Spongy tissue of bone.

Osteoporosis: A disease of the bone leading to an increased risk of fracture.

WHO: World Health Organization.

Diet Monitoring Software

Marion Cottingham
University of Western Australia, Australia

INTRODUCTION

Obesity has been a known problem for over 60 years. As early as 1943, Metropolitan Life Insurance Company declared "Overweight is so common that it constitutes a national health problem of the first order." In 1952, the American Heart association identified obesity as a cardiac risk factor (AHA, 1952). In 1974, obesity was identified as "the most important nutritional disease in the affluent countries of the world" (LANCET editorial, 1974).

Over a few decades, the obesity epidemic has continually been creeping up in all developed countries around the world; this has accelerated rapidly in the last decade, and it appears to have reached a crisis level with unprecedented numbers, particularly in America, joining the overweight or obese categories (Anderson, Konz, Frederich, & Wood, 2001; Mokdad, Serdula, Dietz, Bowman, Marks, Koplan, 1999).

In 2003, the U.S. Surgeon General Richard Carmona said in addressing the House of Representatives, "I welcome this chance to talk with you about a health crisis affecting every State, every city, every community, and every school across our great Nation. The crisis is obesity. It's the fastest growing cause of death in America."

To add to the problem, from 1952 until 2003, the ratio of food prices to all other goods dropped by 12%, making it easier to overindulge in our favorite foods (Philipson, Dai, Helmchen, & Variyam, 2004). In recent years, fresh fruit and vegetables have become more expensive as a proportion of disposable income, while foods with refined grains, sugars, and fats have become cheaper (Sturm, Dai, Helmchen, & Variyam, 2005). Females trying to balance work and family life often turn to the convenience of prepackaged foods with short preparation times, and take-away foods (Lin, Guthrie, & Blaycock, 1996; McCrory, Fuss, Hays, Vinken, Greenberg, & Roberts, 1999; Schluter & Lee, 1999). These tend to be high in calories (Nielsen, Siega-Riz, Popkin, 2002). The environment has become obesogenic (Hill, Wyatt, Reed, & Peters, 2003); Ulijaszek (2007) has called this a "disorder of convenience" (p. 185).

Fast food does not normally equate with healthy food. Certain foods can be high in fat, salt, sugar, and calories. They can also be low in nutrition. Salt in burgers and chips, and phospohorous found in soft drinks, contribute to calcium losses. Sugar causes leaching of calcium, magnesium, and B vitamins. Foods kept warm for long periods experience loss of Vitamin C. Cooking causes losses of B and C vitamins, while fried foods increase our need for the antioxidants A, B, C, E, and zinc.

Weight management has tended to be based on calorie counting (Dietary Guidelines Advisory Committee Report, http://www.health.gov/dietaryguidelines/dgq2005/report/PDF/D2_Energy.pdf) with percentages of calories being split between the three elements: protein, carbohydrates, and fat. But the diet for optimum health also needs to consider the levels of essential vitamins and minerals being consumed. The requirements of all these elements, vitamins and minerals vary with age and gender. This makes it very difficult to monitor your diet without the help of diet monitoring software.

Note: For convenience, the term obesity is used in this article to mean overweight or obese.

BACKGROUND

According to the World Health Organization, more than one billion adults are overweight worldwide, and more than 300 million of these are classed as obese (http://www.who.int/dietphysicalactivity/publications/facts/obesity/en/).

An estimated 17.6 million children under five years old are overweight worldwide; this has doubled since 1980, while the number of overweight adolescents have trebled over the same period.

The body mass index (BMI) measurement is used to identify whether a person is overweight or obese. This is a statistical measure that was developed between 1830 and 1850 by the Belgian polymathematician Adolphe Quetelet. The BMI is still widely in use today. The BMI formula is BMI = weight in kilos, divided by height in meters squared. In general, overweight people are defined as having a BMI over 25, obese people over 30, and morbidly obese people, over 35. BMI measurements tend to increase amongst middle-aged and elderly people, which significantly contribute to osteoarthritis.

Overweight and obesity are defined as having excessive body fat that may impair the health of such individuals. This body fat is normally accumulated over many years of energy imbalance between the calories consumed and the calories expended—in other words, due to overeating and a sedentary lifestyle. A pound of body fat is achieved by consuming 3,500 calories more than the calories being burned off as energy. Even adding one plain biscuit containing 65 calories to your daily diet will make you gain almost 7 pounds per year. Over time, this accumulates into stones. Dieting to lose weight without doing any physical activity causes muscle loss, and can be harmful to an individual's health. Muscle cells contain mitochondria cellular organelles which consume fat. The more muscle you have, the more mitochondria also increases in number, consuming more fat. Consuming less fat causes the mitochondria to use up the fat cells stored in your body as their source of fuel.

Obesity differs between countries with China, Japan, and some African nations at one end of the scale, falling below 5%, while at the other end, the incidence in urban Samoa rises above 75%.

To add to the confusion regarding the number of calories you need to maintain weight, our muscles start to deteriorate by 1% per year by the age of 45 (Sceppa & Layne, 2005). Less muscle lowers the metabolic rate, and that means fewer calories are required at the resting rate. So maintaining exactly the same diet and exercise regime will see individuals overeating without even realizing it. Sceppa's research into reversing muscle loss examines protein deficiency that forces the body to use the amino acids from muscle tissue, leading to muscle loss. The researchers are now reassessing the adult dietary protein requirement, which is currently 0.8 grams per kilo of body weight.

Becoming overweight can adversely effect temperature regulation and put strain on the body's organs and systems over time. This often leads on to high blood pressure, high blood cholesterol and triglycerides, and insulin resistance. Continuing into obesity is known to increase the risk of serious health problems in some individuals, and the more excess fat you have, the higher your risk of chronic disease and disability (Must, Spadano, Coakley, Field, Colditz, & Dietz, 1999). These risks include respiratory difficulties, chronic musculoskeletal problems, skin problems, and infertility at the less serious end, Type 2 diabetes and high blood pressure at the middle of the range, to life-threatening diseases such as some forms of cancer, gallbladder disease, heart disease, and stroke at the most serious end. Other less life-threatening diet related ailments include anemia, ulcers, constipation, irritable bowel syndrome, and kidney stones.

According to the World Health Organization, heart disease and stroke are the world's number one cause of death in the developed world with an estimated 17 million people each year being struck down prematurely (World Health Organization FactSheet, 2006).

Table 1. Recommended dietary allowance of calories for sedentary lifestyle

Age	**Female**	**Male**
0-5 months	108 * weight	108 * weight
5 – 12 months	98 * weight	98 * weight
1 – 3 years	102 * weight	102 * weight
4 – 6 years	90 * weight	90 * weight
7 – 10 years	70 * weight	70 * weight
11 - 14 years	47 * weight	55 * weight
15 – 18 years	40 * weight	45 * weight
19 – 24 years	38 * weight	40 * weight
25 – 50 years	36 * weight	37 * weight
51 and over	30 * weight	30 * weight

Analysis in the World Health Report, 2002 claims that around 58% of diabetes, and 21% of ischaemic heart disease, and between 8 and 42% of certain cancers globally occurred in people with a BMI measurement over 21. So the premise that eating a sound, balanced diet is more likely to lead to a much healthier and longer life has been well proven by medical researchers.

In 1977, the Dietary Goals report recommended food labeling (Senate Select Committee, 1977). This has been adopted in all developed countries.

Knowing the number of calories in food allows you to compare similar food products. But how do you know how many calories you actually need? Table 1 gives the figures to maintain your current weight if you lead a low-activity lifestyle. The figures are based on the RDA data courtesy of National Academies Press (http://www.nap.edu) published reports issued by National Academy of Sciences, National Academy of Engineering, Institute of Medicine, and National Research Council, all operating under charters granted by the Congress of the United States. The weights used in Table 1 are in kilograms. These figures are guidelines only, and there are many other variations. If you monitor your weight and count your calories, you will be able to determine how many calories your body actually requires.

If you do medium activity, then add 400; if you do heavy work, then add 900.

Vitamins and minerals are also required from the food we eat. These are required to maintain health and ward off disease. There is a Recommended Dietary Allowance (RDA) for most vitamins and minerals. The RDA was developed during World War II by Lydia J. Roberts, Hazel K. Stiebeling, and Helen S. Mitchell under the auspices of the National Research Council (Committee on Food and Nutrition, 1941). There was a strong possibility that food would need to be rationed, and the governing bodies needed to know the minimum requirement of each element to maintain health. The original RDAs were introduced in 1941 with a built-in "margin of safety." RDA levels have been adjusted throughout the years since then as new research becomes available. More recent research has shown that taking several vitamins above the RDA may prevent heart disease, cancer, osteoporosis, and other chronic diseases (Giovannucci, Ascherio, Rimm, Stampfer, Colditz, & Willett, 1995; Holick, 2004; Weber, 2001)

Until recently, it was thought that taking multivitamin supplements was not essential for health if you ate a healthy diet. A healthy diet consists of lots of whole grains, fresh fruit and vegetables, nuts, seeds, legumes, unprocessed oils, flesh, and dairy produce should contain an abundance of vitamins and minerals. However, many are lost in the modern processing procedures. Over farming on depleted soils, use of chemical fertilizers, growth hormones, pesticides, artificial ripening, extended storage, milling, heating, cooking, exposure to light and air all have a detrimental effect on vitamins and mineral content. Vitamins contained in fresh fruit and vegetables also deteriorate with time in storage before consumption. Nowadays, it is recommended that most adults take a daily multivitamin as a nutritional safety net. This allows for these variations in the nutrients of food.

To complicate things further there is the protein, carbohydrate, and fat balance in our diet—this is called the PCF ratio. The PCF ratio differs according to your age, height, and muscle mass. Tables 2 to 4 provide the accepted Macronutrient Distribution Ranges (AMDR) for each element's ratio for disease avoidance.

Proteins are broken down by your digestive system to amino acids that help keep your immune system healthy.

Table 2. AMDR protein requirements for each age group

Age	Female	Male
0-6 months	9.1	9.1
7 – 12 months	13.5	13.5
1 – 3 years	13	13
4 – 8 years	19	19
9 – 13 years	34	34
14 - 18 years	46	52
19 – 30 years	46	56
31 and over	46	56

Table 3. AMDR carbohydrate requirements for each age group

Age	Female	Male
0-6 months	60	60
1 – 3 years	130	130
8 and older	130	130

Proteins are also essential to the structure, function, and regulation of the body and help build muscle, which in turn, increases your metabolism. Protein can be found in beef, whole grains, and beans.

Carbohydrates are essential for energy. We tend to think of bread and sugar as being high in carbohydrates, but vegetables and fruit consist mainly of carbohydrates too, and after all, fruit and vegetables are our main source of vitamins and minerals that are so important for good health.

Fats are also known as lipids, and are substances that are insoluble in water. There are three main categories of lipids found in food: cholesterol, saturated fats, and unsaturated fats (poly and mono). All fats are high in calories, and should be eaten sparingly. Unsaturated fats stay liquid at room temperature. They are healthier than saturated fats that are known to increase cholesterol levels and clog arteries.

Mono-unsaturated fat is the healthiest type of fat, and includes olive oil, which remains stable at high temperature, and does not easily become saturated. Foods containing mono-unsaturated fats include almonds, avocado, Brazil nuts, and cashews. Mono-unsaturated fats are believed to lower cholesterol, and thus, your risk of heart disease. Poly-unsaturated fat includes vegetable oils, which provide essential fatty acids for a healthy skin, and for the development of body cells. Polyunsaturates that are rich in omega-3 such as flax oil, walnuts, or oily fish are considered healthy.

Saturated fats are found in whole milk, cheese, butter, lard, and hard margarine. Trans-fatty acid is also a saturated fat. Saturated fats remain solid at room temperature, and are bad for your heart arteries.

Table 4. AMDR fat requirements for each age group

Age	Female	Male
0-6 months	4.9	4.9
7 – 12 months	7.7	7.7
1 – 3 years	7.7	7.7
4 – 8 years	10.9	10.9
9 – 13 years	11	13.2
14 - 18 years	12.1	17.6
19 - 30 years	18.1	18.6
31 – 50 years	15.1	18.6
51 and over	15.1	18.6

Anyone who has ever been in a diet knows how tedious it is just to count the calories they are consuming. If you also had to ensure the PCF ratio was adhered to, and that you were getting the correct amount of 33 different vitamin and minerals—monitoring your diet would be a full-time job. The Balance Your Diet monitoring software described in this article does all that for you.

BALANCE YOUR DIET: DIET MONITORING SOFTWARE

The Balance Your Diet software application described in this article is for people who want to maintain the ultimate balanced diet for good general health, longevity, and well-being. Balance Your Diet allows you to add the food items consumed throughout your day. The information about the food items come from an adapted version of the USDA food database that contains over 6,000 food products, each having information on 33 different nutrients the product contains. The person using the software chooses the products they consume; these are added to the Food Product Description list, which is found under the date at the left hand side of the Main screen (see Figure 1). This makes it easier to select food consumed during the current day, as most people only eat from a few hundred different products.

Balance Your Diet also calculates the recommended daily allowances for the age and gender provided by the dieter, and displays these amounts in the rightmost column. This allows the dieter to know what to aim for.

Figure 2 shows the adapted USDA database from Balance Your Diet. The food is kept in their prospective categories, which are displayed as option buttons along the top. This makes products easier to find. There is also a search engine that will list all products from the whole database that contain the search word (see Figure 3).

Balance Your Diet is the ultimate diet tool, not just for weight maintenance, but for anyone who is concerned about whether or not their diet is healthy. By comparing the nutrients consumed with the recommended daily allowances of all the nutritional components that are essential for health, it can alert you to deficiencies in any vitamin or mineral. It also keeps track of your protein–carbohydrate–fat ratio for the particular diet you are following.

Figure 1. The RDA amounts for calories, main elements, and minerals that optimize your diet for health

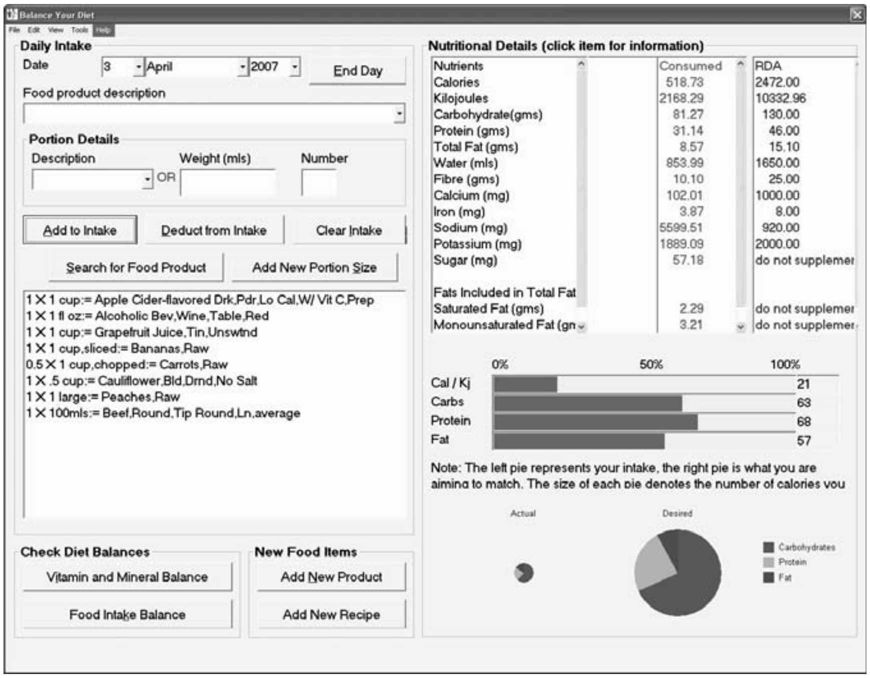
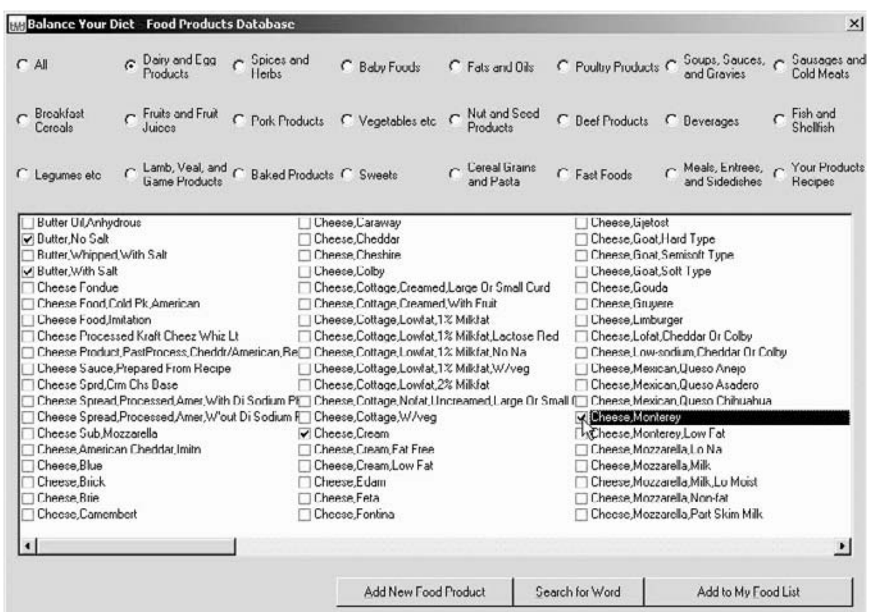


Figure 2. Adapted USDA database showing the food groups available and some of the products available when you choose dairy and egg products



D

Figure 3. Searching for a word from a food product displays any product from the full database that contains that word

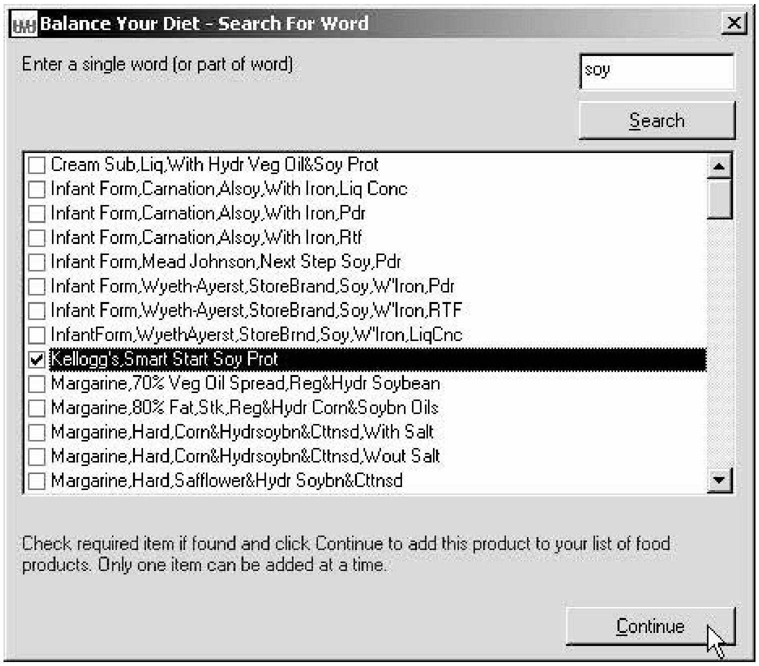


Figure 4. Selecting your frame type will display your ideal weight according to your personal details

Balance Your Diet will also calculate your BMI information. Table 5 shows the BMI figures color-coded from underweight to obese. The top table uses meters and kilograms, while the bottom table uses feet and pounds.

The tables show that there is a range of BMIs in each category. For instance, people with a height of 1.60 meters are in the ideal category, if their weight lies between 50 kilograms (BMI 19.5) and 60 kilograms (BMI 23.4).

Balance Your Diet adjusts the BMI further by taking a person's frame type into account. Entering a wrist or elbow measurement displays the appropriate frame type. If you know your frame type, you can select it from a drop-down box, and immediately, the dialog will be updated to display the ideal weight for your height and gender (see Figure 4).

Knowing your height, weight, and frame type Balance Your Diet can calculate your exact BMI and ideal weight range. Figure 5 shows the information displaying this information. The frame type can be calculated by entering your wrist or elbow measurement.

Balance Your Diet provides the number of calories required to maintain the current weight, based on the RDA. It also allows a goal weight to be entered, and

Table 5. BMI numbers color coded for underweight, ideal, overweight, and obese

Body Mass Index (BMI)

Height																				
2.00	7.5	8.8	10.0	11.3	12.5	13.8	15.0	16.3	17.5	18.8	20.0	21.3	22.5	23.8	25.0	26.3	27.5	28.8	30.0	31.3
1.95	7.9	9.2	10.5	11.8	13.1	14.5	15.8	17.1	18.4	19.7	21.0	22.4	23.7	25.0	26.3	27.6	28.9	30.2	31.6	32.9
1.90	8.3	9.7	11.1	12.5	13.9	15.2	16.6	18.0	19.4	20.8	22.2	23.5	24.9	26.3	27.7	29.1	30.5	31.9	33.2	34.6
1.85	8.8	10.2	11.7	13.1	14.6	16.1	17.5	19.0	20.5	21.9	23.4	24.8	26.3	27.8	29.2	30.7	32.1	33.6	35.1	36.5
1.80	9.3	10.8	12.3	13.9	15.4	17.0	18.5	20.1	21.6	23.1	24.7	26.2	27.8	29.3	30.9	32.4	34.0	35.5	37.0	38.6
1.75	9.8	11.4	13.1	14.7	16.3	18.0	19.6	21.2	22.9	24.5	26.1	27.8	29.4	31.0	32.7	34.3	35.9	37.6	39.2	40.8
1.70	10.4	12.1	13.8	15.6	17.3	19.0	20.8	22.5	24.2	26.0	27.7	29.4	31.1	32.9	34.6	36.3	38.1	39.8	41.5	43.3
1.65	11.0	12.9	14.7	16.5	18.4	20.2	22.0	23.9	25.7	27.5	29.4	31.2	33.1	34.9	36.7	38.6	40.4	42.2	44.1	45.9
1.60	11.7	13.7	15.6	17.6	19.5	21.5	23.4	25.4	27.3	29.3	31.3	33.2	35.2	37.1	39.1	41.0	43.0	44.9	46.9	48.8
1.55	12.5	14.6	16.6	18.7	20.8	22.9	25.0	27.1	29.1	31.2	33.3	35.4	37.5	39.5	41.6	43.7	45.8	47.9	49.9	52.0
1.50	13.3	15.6	17.8	20.0	22.2	24.4	26.7	28.9	31.1	33.3	35.6	37.8	40.0	42.2	44.4	46.7	48.9	51.1	53.3	55.6
1.45	14.3	16.6	19.0	21.4	23.8	26.2	28.5	30.9	33.3	35.7	38.0	40.4	42.8	45.2	47.6	49.9	52.3	54.7	57.1	59.5
	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100	105	110	115	120	125

Kilograms

Directions: Find your weight along the bottom of the chart, and move straight upwards until you come to the row that matches your height. Green marks ideal BMIs, while orange marks overweight, red obese, and yellow underweight.

Body Mass Index (BMI)

Height																						
6'6"	8.1	9.2	10.4	11.6	12.7	13.9	15.0	16.2	17.3	18.5	19.6	20.8	22.0	23.1	24.3	25.4	26.6	27.7	28.9	30.0	31.2	32.4
6'5"	8.3	9.5	10.7	11.9	13.0	14.2	15.4	16.6	17.8	19.0	20.2	21.3	22.5	23.7	24.9	26.1	27.3	28.5	29.6	30.8	32.0	33.2
6'4"	8.5	9.7	11.0	12.2	13.4	14.6	15.8	17.0	18.3	19.5	20.7	21.9	23.1	24.3	25.6	26.8	28.0	29.2	30.4	31.6	32.9	34.1
6'3"	8.7	10.0	11.2	12.5	13.7	15.0	16.2	17.5	18.7	20.0	21.2	22.5	23.7	25.0	26.2	27.5	28.7	30.0	31.2	32.5	33.7	35.0
6'2"	9.0	10.3	11.6	12.8	14.1	15.4	16.7	18.0	19.3	20.5	21.8	23.1	24.4	25.7	27.0	28.2	29.5	30.8	32.1	33.4	34.7	35.9
6'1"	9.2	10.6	11.9	13.2	14.5	15.8	17.2	18.5	19.8	21.1	22.4	23.7	25.1	26.4	27.7	29.0	30.3	31.7	33.0	34.3	35.6	36.9
6'0"	9.5	10.8	12.2	13.6	14.9	16.3	17.6	19.0	20.3	21.7	23.1	24.4	25.8	27.1	28.5	29.8	31.2	32.5	33.9	35.3	36.6	38.0
5'11"	9.8	11.2	12.6	13.9	15.3	16.7	18.1	19.5	20.9	22.3	23.7	25.1	26.5	27.9	29.3	30.7	32.1	33.5	34.9	36.3	37.7	39.1
5'10"	10.0	11.5	12.9	14.3	15.8	17.2	18.7	20.1	21.5	23.0	24.4	25.8	27.3	28.7	30.1	31.6	33.0	34.4	35.9	37.3	38.7	40.2
5'9"	10.3	11.8	13.3	14.8	16.2	17.7	19.2	20.7	22.2	23.6	25.1	26.6	28.1	29.5	31.0	32.5	34.0	35.4	36.9	38.4	39.9	41.3
5'8"	10.6	12.2	13.7	15.2	16.7	18.2	19.8	21.3	22.8	24.3	25.8	27.4	28.9	30.4	31.9	33.5	35.0	36.5	38.0	39.5	41.1	42.6
5'7"	11.0	12.5	14.1	15.7	17.2	18.8	20.4	21.9	23.5	25.1	26.6	28.2	29.8	31.3	32.9	34.5	36.0	37.6	39.2	40.7	42.3	43.9
5'6"	11.3	12.9	14.5	16.1	17.8	19.4	21.0	22.6	24.2	25.8	27.4	29.1	30.7	32.3	33.9	35.5	37.1	38.7	40.4	42.0	43.6	45.2
5'5"	11.6	13.3	15.0	16.6	18.3	20.0	21.6	23.3	25.0	26.6	28.3	30.0	31.6	33.3	34.9	36.6	38.3	39.9	41.6	43.3	44.9	46.6
5'4"	12.0	13.7	15.4	17.2	18.9	20.6	22.3	24.0	25.7	27.5	29.2	30.9	32.6	34.3	36.0	37.8	39.5	41.2	42.9	44.6	46.3	48.1
5'3"	12.4	14.2	15.9	17.7	19.5	21.3	23.0	24.8	26.6	28.3	30.1	31.9	33.7	35.4	37.2	39.0	40.7	42.5	44.3	46.1	47.8	49.6
5'2"	12.8	14.6	16.5	18.3	20.1	21.9	23.8	25.6	27.4	29.3	31.1	32.9	34.8	36.6	38.4	40.2	42.1	43.9	45.7	47.6	49.4	51.2
5'1"	13.2	15.1	17.0	18.9	20.8	22.7	24.6	26.5	28.3	30.2	32.1	34.0	35.9	37.8	39.7	41.6	43.5	45.3	47.2	49.1	51.0	52.9
5'0"	13.7	15.6	17.6	19.5	21.5	23.4	25.4	27.3	29.3	31.2	33.2	35.2	37.1	39.1	41.0	43.0	44.9	46.9	48.8	50.8	52.7	54.7
4'11"	14.1	16.2	18.2	20.2	22.2	24.2	26.3	28.3	30.3	32.3	34.3	36.4	38.4	40.4	42.4	44.4	46.5	48.5	50.5	52.5	54.5	56.6
	70	80	90	100	110	120	130	140	150	160	170	180	190	200	210	220	230	240	250	260	270	280

Pounds

Directions: Find your weight along the bottom of the chart, and move straight upwards until you come to the row that matches your height. Green marks ideal BMIs, while orange marks overweight, red obese, and yellow underweight.

the date the goal is to be achieved. It calculates the calories per day to achieve this, and the person can choose to work towards this by choosing to display the goal weight calories in the main screen. Figure 6 shows the Personal Details screen with both the RDA and the goal calorie numbers on display.

Balance Your Diet calculates the RDA vitamins and minerals required, according to personal details provided (see Figure 6). Figure 7 shows the vitamin and mineral intake over one day's consumption. The aim is to achieve a bar that is further than the first RDA area. The table shows the amounts actually consumed compared with the RDA recommendations. The last column gives the maximum safe levels to give the person reassurance that they are not going to come to any harm.

Figure 5. BMI information based on the personal information provided

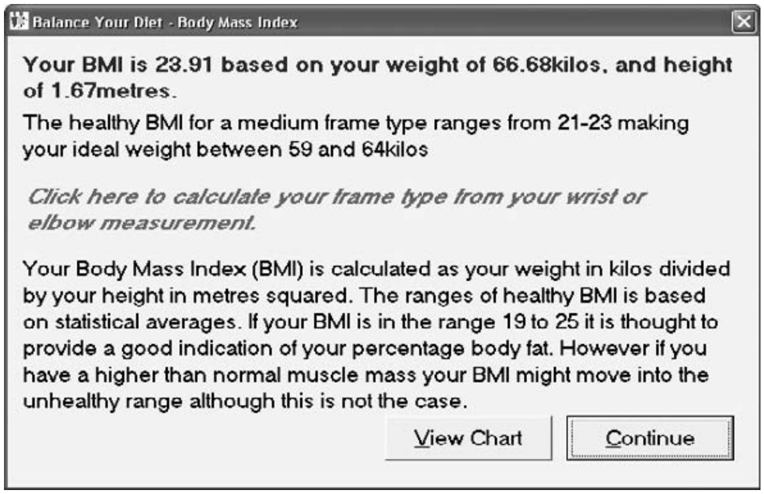


Figure 6. Personal Detail screen showing the RDA calorie requirements and the goal weight by goal date calorie requirements

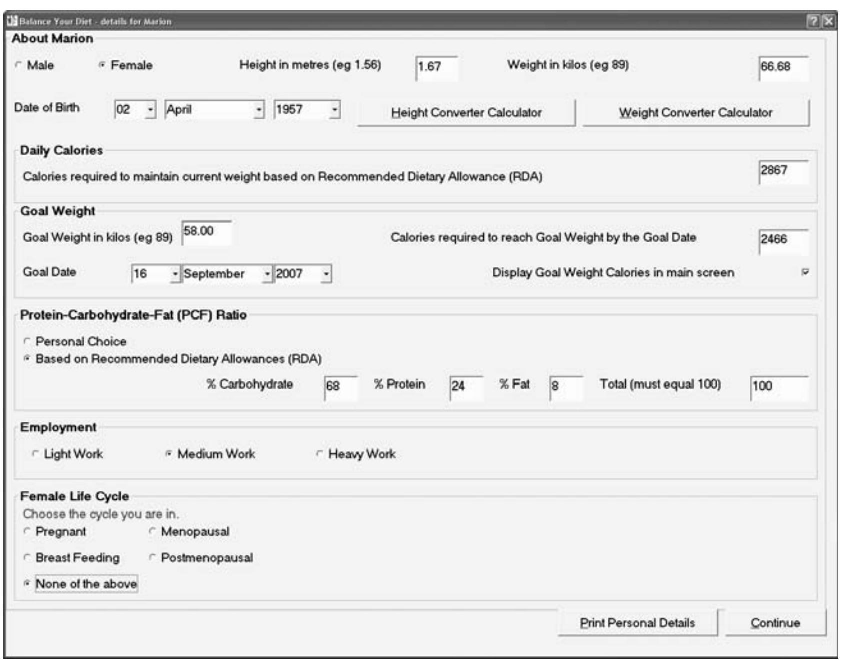
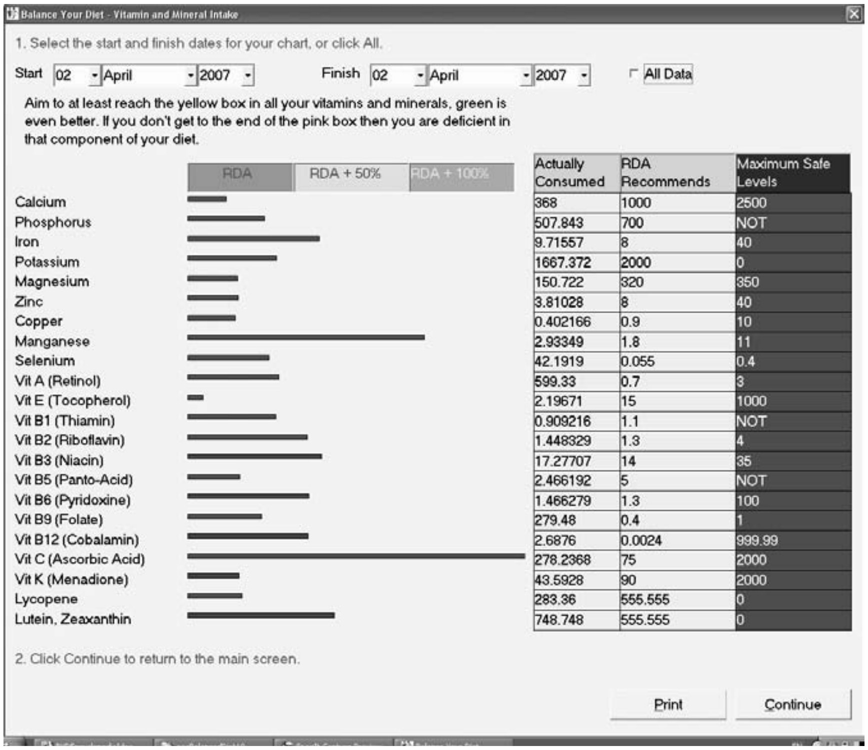
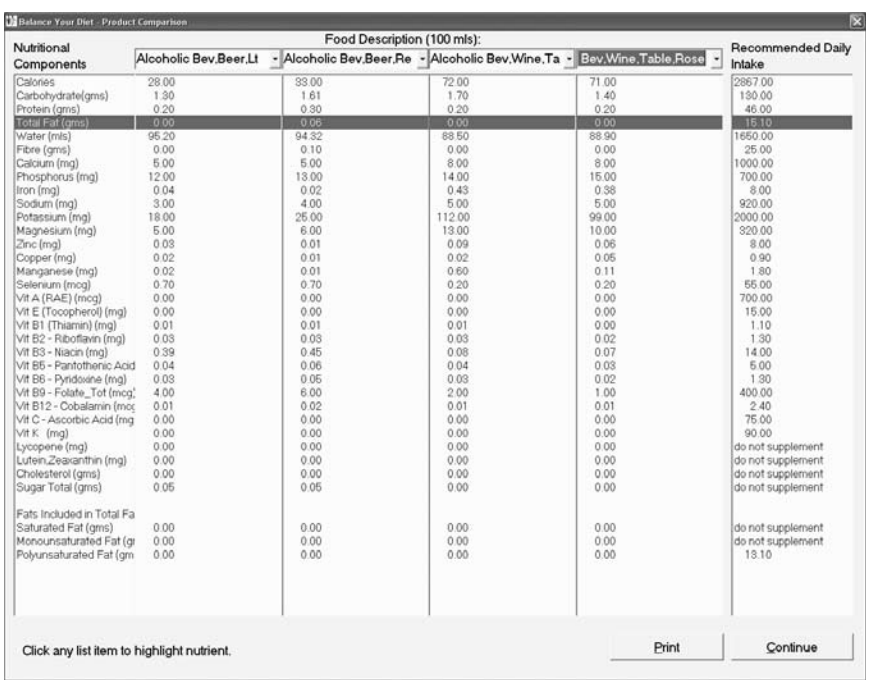


Figure 7. Vitamin and mineral intake screen shows the actual amounts consumed in comparison with the desired amounts; it also shows the maximum safety levels



	Actually Consumed	RDA Recommends	Maximum Safe Levels
Calcium	368	1000	2500
Phosphorus	507.843	700	NOT
Iron	9.71557	8	40
Potassium	1667.372	2000	0
Magnesium	150.722	320	350
Zinc	3.81028	8	40
Copper	0.402166	0.9	10
Manganese	2.93349	1.8	11
Selenium	42.1919	0.055	0.4
Vit A (Retinol)	599.33	0.7	3
Vit E (Tocopherol)	2.19671	15	1000
Vit B1 (Thiamin)	0.909216	1.1	NOT
Vit B2 (Riboflavin)	1.448329	1.3	4
Vit B3 (Niacin)	17.27707	14	35
Vit B5 (Panto-Acid)	2.466192	5	NOT
Vit B6 (Pyridoxine)	1.466279	1.3	100
Vit B9 (Folate)	279.48	0.4	1
Vit B12 (Cobalamin)	2.6876	0.0024	999.99
Vit C (Ascorbic Acid)	278.2368	75	2000
Vit K (Menadione)	43.5928	90	2000
Lycopene	283.36	555.555	0
Lutein, Zeaxanthin	748.748	555.555	0

Figure 8. Comparing the nutrient values from four similar food products



Nutritional Components	Alcoholic Bev,Beer,Lt	Alcoholic Bev,Beer,Re	Alcoholic Bev,Wine,Ta	Bev,Wine,Table,Rose	Recommended Daily Intake
Calories	28.00	33.00	72.00	71.00	2867.00
Carbohydrate(gms)	1.30	1.61	1.70	1.40	130.00
Protein (gms)	0.20	0.30	0.20	0.20	46.00
Total Fat (gms)	0.00	0.06	0.00	0.00	15.10
Water (mls)	95.20	94.32	88.50	88.90	1650.00
Fibre (gms)	0.00	0.10	0.00	0.00	25.00
Calcium (mg)	5.00	5.00	8.00	8.00	1000.00
Phosphorus (mg)	12.00	13.00	14.00	15.00	700.00
Iron (mg)	0.04	0.02	0.43	0.38	8.00
Sodium (mg)	3.00	4.00	5.00	5.00	920.00
Potassium (mg)	18.00	25.00	112.00	99.00	2000.00
Magnesium (mg)	5.00	6.00	13.00	10.00	320.00
Zinc (mg)	0.03	0.01	0.09	0.06	8.00
Copper (mg)	0.02	0.01	0.02	0.05	0.90
Manganese (mg)	0.02	0.01	0.60	0.11	1.80
Selenium (mcg)	0.70	0.70	0.20	0.20	55.00
Vit A (RAE) (mcg)	0.00	0.00	0.00	0.00	700.00
Vit E (Tocopherol) (mg)	0.00	0.00	0.00	0.00	15.00
Vit B1 (Thiamin) (mg)	0.01	0.01	0.01	0.00	1.10
Vit B2 - Riboflavin (mg)	0.03	0.03	0.03	0.02	1.30
Vit B3 - Niacin (mg)	0.39	0.45	0.08	0.07	14.00
Vit B5 - Pantothenic Acid	0.04	0.06	0.04	0.03	5.00
Vit B6 - Pyridoxine (mg)	0.03	0.05	0.03	0.02	1.30
Vit B9 - Folate_Tot (mcg)	4.00	6.00	2.00	1.00	400.00
Vit B12 - Cobalamin (mcg	0.01	0.02	0.01	0.01	2.40
Vit C - Ascorbic Acid (mg	0.00	0.00	0.00	0.00	75.00
Vit K (mg)	0.00	0.00	0.00	0.00	90.00
Lycopene (mg)	0.00	0.00	0.00	0.00	do not supplement
Lutein,Zeaxanthin (mg)	0.00	0.00	0.00	0.00	do not supplement
Cholesterol (gms)	0.00	0.00	0.00	0.00	do not supplement
Sugar Total (gms)	0.05	0.05	0.00	0.00	do not supplement
Fats Included in Total Fa					
Saturated Fat (gms)	0.00	0.00	0.00	0.00	do not supplement
Monounsaturated Fat (gr	0.00	0.00	0.00	0.00	do not supplement
Polyunsaturated Fat (gm	0.00	0.00	0.00	0.00	13.10

Figure 9. Foods can be ordered according to a specific nutrient; this figure shows foods being ordered according to their levels of calcium

Figure 10. Pie charts showing the actual consumption compared with the desired consumption

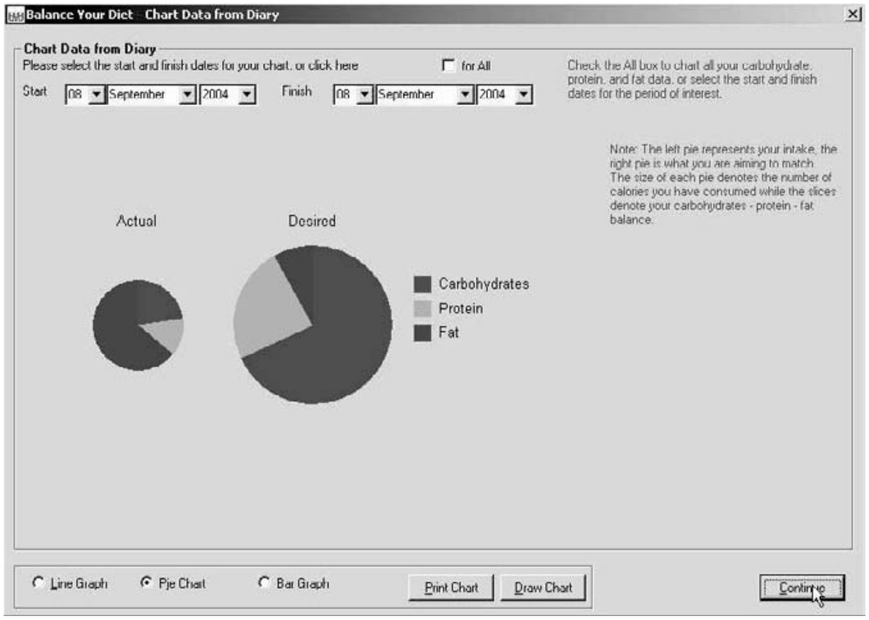


Table 6. Popular diets and their PCF ratios

Diet Name	Protein %	Carbohydrate %	Fat %
National Research Council US	15	55	30
California Diet	18	54	28
Pritiken Program	12.5	80	7.5
Atkins	46	13	15.1

Figure 8 shows the nutrients listed for four different food products. This allows you to pick the one that gets you the closest to your optimal diet.

If you are unsure what foods contain a particular nutrient, you can sort foods from the whole database in high-to-low or low-to-high order. Figure 9 shows foods containing calcium, ordered from high-to-low.

Figure 10 shows two pie charts. The pie on the left represents the food being consumed that day. The pie on the right is the desired result by the end of the day. The slices represent the three PCF ratios, and the size of the pie represents the calories consumed.

Different diets usually are variations of the PCF ratios. Table 6 gives four popular diets. It appears that if you reduce one of the components you lose weight. Whether or not it is safe to do so is another discussion.

FUTURE TRENDS

There is no quick solution to the obesity epidemic. Dieting appears to be something that most people rebound from, and end up bigger (and worse off) than they were when they started. How do you stop people from eating high calorie food when it is more convenient to serve up? How do you get people to walk when they have a car? How do you get people to climb stairs when there is an elevator? How do you get people to change their habits? How do you educate people about food?

People do not often react until they have some severe medical problem, but by then, it is too late. For example, if people are at a high risk of Type 2 diabetes, surely it makes sense to go on a sensible diet now, rather than wait for the inevitable. After the diagnosis, their diet is going to be restricted anyway, and they will be denied the odd binge. The message to obese people should be that prevention is better than cure.

Governments are looking at extra taxes on food manufacturers based on the amount of fat they are including in their products—a fat tax. They are also considering restricting advertisements aimed at children. But the supermarkets are still displaying candy and potato chips at the checkouts, nice and low so that little prying eyes cannot miss. These are also tempting for the adults who are stopped there, waiting to be served.

CONCLUSION

BYD Software is designed to help you achieve a balanced and healthy diet. Doing the research for this software package, we came to understand how complicated it is to fully understand what is healthy for you. Balance Your Diet allows you to:

- Count calories as part of your weight loss program
- Monitor vitamin and minerals in food products for conscious nutrition
- Monitor the PCF ratio in your diet

However, the first step in overcoming the obesity problem is to convince overweight and obese people that they need to go on a diet. The second step is to get them exercising. The third step (and the hardest) is to change their old habits, and keep them on the straight and narrow path to a healthier existence. It is all-too-easy to slip back into the sedentary lifestyle with the take-away food.

D

REFERENCES

AHA, Harvard School of Public Health, Department of Nutrition. (1952). *Food for your heart: A manual for patient and physician. New* York: American Heart Association.

Anderson, J., Konz, E., Frederich, R., & Wood, C. (2001). Long term weight-loss maintenance: A meta-analysis of U.S. studies. *American Journal of Clinical Nutrition, 74*, 579–584.

Committee on Food and Nutrition. (1941). *Recommended dietary allowances*. 1941 Nutrition Division, Federal Security Agency Washington, D.C.

Giovannucci, E., Ascherio, A., Rimm, E.B., Stampfer, M.J., Colditz, G.A., & Willett, W.C. (1995). Intake of caretenoids and retinal in relation to risk of prostrate cancer. *Jol Natl Cancer Inst., 87*, 1767–1776.

Hill, J.O., Wyatt, H.R., Reed G.W., & Peters, J.C. (2003). Obesity and the environment where do we go from here? *Science, 299*, 853–835.

Holick, M.F. (2004). Vitamin D: importance in the prevention of cancers, type 1 diabetes, heart disease, and osteoporosis. *Amer Jnl Clinical Nutrition, 79*, 362–371.

LANCET Editorial. (1974). Infant and adult obesity. I:17–18.

Lin, B., Guthrie, J., & Blaycock, J.R. (1996). The diets of America's children: Influences on dining out, household characteristics, and nutritional knowledge. *Agric Econ Rep, 746*, 1–37.

McCrory, M.A., Fuss, P.J., Hays, N.P., Vinken, A.G., Greenberg, A.S., & Roberts, S.B. (1999). Overeating in America: Association between restaurant food consumption and body fatness in healthy adult men and women ages 19–80. *Obes Res, 7*, 564–571.

Mokdad, A.H., Serdula, M.K., Dietz, W.H., Bowman, B.A., Marks, I.S., & Koplan, J.P. (1999). The spread of the obesity epidemic in the United States, 1991–1998. *JAMA, 282*, 1519–1522.

Must, A., Spadano, J., Coakley, E.H., Field, A.E., Colditz, G., & Dietz, W.H. (1999). The disease burden associated with overweight and obesity. *JAMA, 282*, 1523–1529.

Nielsen, S.J., Siega-Riz, A.M., & Popkin, B.M. (2002). Trends in energy intake in U.S. between 1977 and 1996: Similar shifts seen across age groups. *Obes Res, 10*, 370–378.

Philipson, T., Dai, C., Helmchen, L., Variyam, J.N. (2004). *The economics of obesity.* A report on the workshop held at USDA's Economic Research Service. E-FAN-04-004, USDA/ERS.

Sceppa, C.C., & Layne, J. (2005). Low Protein + Low Exercise = Sarcopenia. *Agriculture Research Magazine, 53*(5), 14–16.

Schluter, G., & Lee, C. (1999). Changing good consumption patterns: Their effects on the U.S. Food System, 1972–92. *Food Rev, 22*, 35–37.

Senate Select Committee on Nutrition and Human Needs (U.S.). (1977). *Dietary goals for the United States, 2nd Ed.* Washington: Government Printing Office.

Sturm, R. (2005). Childhood obesity: What we can learn from existing data on societal trends, Part 2. *Prev Chronic Dis, 2*, A20.

Ulijaszek, S.J. (2007). Obesity: A disorder of convenience. In *The International Association for the Study of Obesity: Obesity Reviews 8* (Suppl 1) (pp. pp183–187).

Weber, P. (2001). Vitamin K and bone health. Nutrition 17:880–887.

World Health Organization. (2006). *Obesity and overweight.* Fact Sheet No.311. Retrieved February 4, 2008, from http://www.who.int/mediacentre/factsheet/fs311

KEY TERMS

Body Mass Index (BMI): A measurement used to identify whether a person is overweight or obese.

Chronic Disease Risk: Likelihood of contracting a chronic disease, such as diabetes.

Diet: Total combination of food intake.

Minerals: Naturally occurring substance required by living organisms for a healthy diet.

Nutrition: A science to explain the metabolic and physiological responses of the body to diet.

Obesity: Excessively overweight.

Overweight: Above average weight.

PCF: Ratio of protein, carbohydrates, and fats.

Vitamins: An organic compound required by living organisms.

D

Distributed Medical Image and Volume Registration

Roger Tait
Nottingham Trent University, UK

Gerald Schaefer
Aston University, UK

INTRODUCTION

The ability to visualise hidden structures in detail using 3-D volume data has become a valuable resource in medical imaging applications (Maintz & Viergever, 1998). Importantly, the alignment of volumes enables the combination of different structural and functional information for diagnosis and planning purposes (Pluim, Maintz, & Viergever, 2003). Transform optimisation, resampling, and similarity calculation form the basic stages of a registration process (Zitova & Flusser, 2003): During transform optimisation, translation and rotation parameters which geometrically map points in the reference (fixed) image/volume to points in the sensed (moving) image/volume are estimated. Once estimated, pixel/voxel intensities which are mapped into nondiscrete coordinates are interpolated during the resampling stage. After resampling, a metric is used for similarity calculation in which the degree of likeness between corresponding volumes is evaluated (Tait & Schaefer, 2008). Optimisation of the similarity measure is the goal of the registration process and is achieved by seeking the best transform. All possible transform parameters therefore define the search space. Due to the iterative nature of registration algorithms, similarity calculation represents a considerable performance bottleneck which limits the speed of time critical clinical applications.

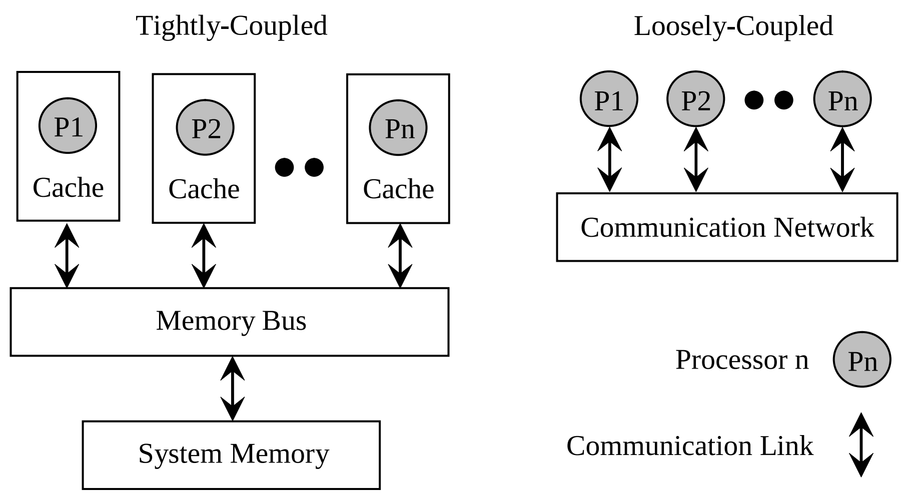
The use of parallel computing to overcome these time constraints has become an important research area (Nicolescu & Jonker, 2000). Conveniently, many of the similarity calculation strategies employed in medical registration are inherently parallel and therefore well suited to distribution. An important consideration when adopting a parallel processing approach is the architecture of the host system. In a computer constructed of multiple processors with shared-memory, data distribution is not required. These systems are viewed as tightly-coupled architectures. In contrast, a loosely-coupled architecture consists of multiple computers in different locations. Loosely-coupled architectures therefore require data distribution, communication, and accumulation mechanisms. Importantly, the most effective distribution scheme will depend on the architecture of the host system (Seinstra, Koelma, & Geusebroek, 2002). The two contrasting architectures of host systems are illustrated in Figure 1.

BACKGROUND

In the context of parallel processing, registration of medical data has been achieved by Warfield, Jolesz, and Kikinis (1998) who introduced a nonrigid algorithm based on the work-pile paradigm. Their goal was to develop an interpatient registration algorithm which can be applied without operator intervention to a database of several hundred scans. In an initial step, each scan is segmented using a statistical classification method. This preprocessing stage is used to identify different tissue types including skin, white matter, grey matter, and bone structure. Once segmented, a transform which brings these features into alignment is estimated. The system employs a message passing interface and cluster of symmetric multiprocessors to execute parallel similarity calculation operations using multiple threads. Crucially, work is dynamically load balanced. Results published by the group show that successful registration of 256 × 256 × 52 volume brain scans can been achieved in minutes rather than hours.

Christensen (1998) compares two nonthreaded architectures, Multiple Instruction Multiple Data (MIMD) and Single Instruction Multiple Data (SIMD). The work presented raises implementation issues and timing analysis for the registration of 32 × 32 × 25, 64 × 64 × 50 and 128 × 128 × 100 volume datasets. During

Figure 1. Tightly vs. loosely-coupled architectures. Data is either fetched from main memory via a memory bus, or is transferred over a communications network.



each clock cycle, the SIMD implementation performs calculations in which all processors are performing the same operation. The MIMD implementation, in contrast, breaks an algorithm into independent parts which are solved simultaneously by separate processors. The movement of data in both shared-memory systems is unrestricted and during execution each processor has access to the whole memory. The main performance bottleneck associated with both approaches was reported as scalability of hardware with increasing numbers of processors. Crucially, the MIMD implementation is recorded as being approximately four times faster than its SIMD counterpart. Reduced performance of the SIMD implementation is reportedly caused by overheads during serial portions of the registration algorithm.

More recently, the demands placed on registration algorithms when aligning deformable structures in 3-D space have been discussed. Salomon, Heitz, Perrin, and Armspach (2005) introduce deformable registration of volumes which involves optimisation of several thousand parameters and typically requires several hours processing on a standard workstation. Based on simulation of stochastic differential equations and using simulated annealing, a parallel approach that yields processing times compatible with clinical routines is presented. The approach represents a hierarchical displacement vector field which is estimated by means of an energy function. The energy function is scaled in relation to the similarly measure and is re-evaluated at the end of each transform parameter optimisation cycle. The algorithm is reportedly suited to massively parallel implementation and has been successfully applied to the registration of 256 × 256 × 256 volumes. Again, the results published demonstrate how alignment can be achieved in minutes rather than hours.

In the next section, a coarse-grained approach to parallelism is described which increases flexibility and allows the issues of fine-grained parallelism to be ignored. Building on a distributed blackboard architecture, the approach adopted supports multiple distributed agents organised in a worker/manager model. Crucially, the basic alignment steps are allocated to individual processors, the most computationally intensive of which are performed concurrently.

SIMILARITY CALCULATION USING A DISTRIBUTED BLACKBOARD ARCHITECTURE

Formally the inputs to a volume-based registration process can be defined as the fixed volume, the moving volume, and the transform used to map voxel coordinates. The goal of the registration process is recovery of a spatial mapping that brings the two volumes into alignment (Yoo, 2004). To achieve this, a metric is employed to generate a measure of similarity based on how well aligned the transformed moving volume is with the fixed volume. The measure of similarity produced forms a quantitative criterion which can be optimised in a search space defined by transform parameters (Penney, Weese, Little, Desmedt, Hill, & Hawkes, 1998). Importantly, by employing a gradient-descend optimisation technique the metric can be used to produce derivatives of the similarity measure with respect to each transform parameter. Using the resulting derivatives, updated transform parameters can be estimated and evaluated.

To parallelise the registration process, both fixed and moving volumes require division into segments and distribution. The similarities between corresponding volume segments can then be estimated concurrently. To achieve this, transform parameters are propagated to all processing nodes in the distributed processing network. On receiving transform parameters, each node computes derivatives of the similarity measure for the segments allocated to it. The local derivatives computed are then accumulated and summed into a global derivative. This allows the transform parameters to be updated based upon the similarity of complete volumes. Convergence testing is then performed using the newly updated transform parameters. Depending on the success or failure of convergence testing, propagation of updated transform parameters and hence evaluation of new parameters can occur.

The iDARBS Framework

Distributed Algorithmic and Rule-based Blackboard System (DARBS) is a distributed blackboard architecture based on a client/server model (Nolle, Wong, & Hopgood, 2001). In DARBS, the server functions as a blackboard and client modules as agents. In general, each agent represents a structure in which rules and algorithms can be embodied. Imaging DARBS (iDARBS), an underlying framework (Tait, Schaefer, & Hopgood, 2006a) on which registration algorithms can be hosted (Tait, Schaefer, & Hopgood, 2006b), consists of Distributor, Worker, and Manager Agent types:

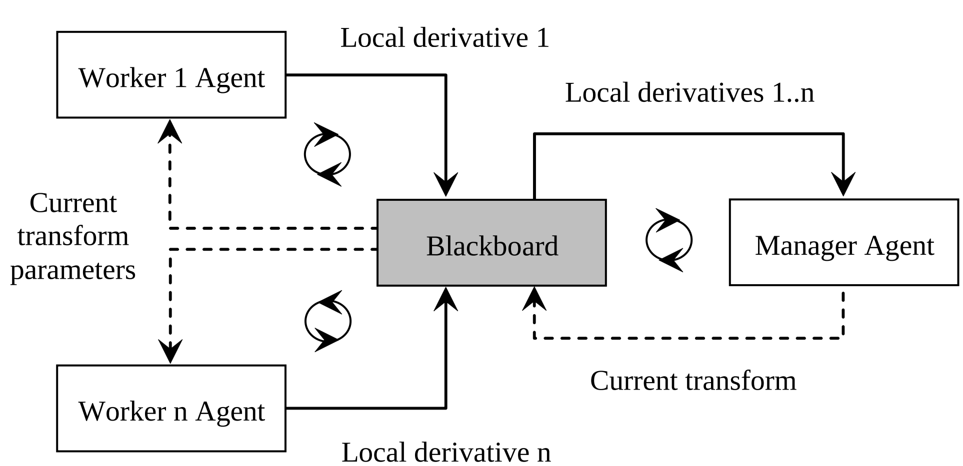
- The Distributor Agent splits selected volumes into segments which are placed on the blackboard.
- Worker Agents take segments from the blackboard and perform local processing.
- The Manager Agent supervises Worker Agent activities.

The accumulation of local derivatives by the Manager Agent and the propagation of updated transform parameters to Worker Agents are illustrated in Figure 2.

Initialisation of the Alignment Process

The Distributor Agent places predefined parameters associated with the optimisation process on the black-

Figure 2. The flow of local derivatives and current transform parameters between framework components



board. The parameters include an initial step length, a minimum step length, and the maximum number of iterations to be performed. Logically, the initial step length is used to initialise the optimisation process. Convergence of optimisation and hence the selection of final transform parameters is controlled by the minimum step length.

To start the alignment process, the Distributor Agent calculates an initial centre of rotation and a translation. Employing both fixed and moving volumes, initial parameters are estimated using centres of mass computed from intensity levels. Once estimated, the fixed volume centre is set as the rotational centre of the initial transform. The translation component, in contrast, is set as the vector between the fixed and moving volume centres. To simplify calculation, no rotation is specified in the initial transform. Crucially, the use of centres of mass over geometrical volume centres results in a robust estimate of the initial transform.

Estimation of Local Derivatives

For similarity between segments to be calculated, the Worker *n* Agent employs a region of interest and current transform parameters retrieved from the blackboard. First, the region of interest is used to identify the fixed segment without borders. Then, for all voxel coordinates within the fixed segment, corresponding moving segment coordinates are computed using the current transform. If the transformation of fixed segment voxel coordinates results in a corresponding location that falls inside the moving segment, the number of valid voxels is incremented and a contribution to the local derivative is made. Otherwise the voxel is considered invalid and the next fixed segment voxel coordinates are processed. By evaluating all voxels, local derivatives of the similarity measure with respect to each transform parameter are generated.

The contribution to a local derivative represents a summation of intensities, from a moving segment gradient image, around the mapped voxel coordinates. Using a recursive Gaussian filter, a gradient image is created from the moving segment. The gradient image represents a vector field in which every vector points in the direction of its nearest edge, an edge being a rapid increase or decrease in neighbouring intensities. Each vector has a magnitude proportional to the second derivative of the intensity in the direction of the vector. Created once after retrieval of the moving segment, the gradient image is used for all iterations of the optimisation process.

Advancing the Transform Parameters

To update the current transform parameters, a gradient-descent optimisation scheme is employed to advance the current transform in the direction of the global derivative. If the direction of the global derivative changes abruptly, it is assumed that an optimum has been encountered and the step length is reduced by a half. Once the step length becomes smaller than the predefined minimum, the optimisation process is considered as having converged. This allows the precision of final transform parameters to be specified. If optimisation of the transform parameters fails to reach the desired precision, the maximum number of iterations is used to halt the optimisation process. Understandably, large numbers of iterations and long computational times result when the initial step length chosen is small. Large step lengths, in contrast, may result in the optimum transform parameters being missed.

EXPERIMENTAL TESTING

Efficiency tests are presented and demonstrate the performance increase of volume registration in nondistributed and distributed processing environments. To determine the accuracy of alignment between volumes, a normalised correlation similarity metric was selected for testing purposes. Timing of an experiment started when the Manager Agent propagates transform parameters to all Worker Agents. Timing stopped when the Manager Agent places the final transform parameters on the blackboard. A sequential algorithm, constructed of the same components, was used as a performance benchmark for comparison. The datasets used for testing, two MRI T1 volumes of an anatomically normal male head (Stark & Bradley, 1999), contain $180 \times 218 \times 180$ voxels with a known translation and rotation. The volume pair was registered three times and average processing times calculated.

Computed using all voxels, normalised correlation calculates the cross-correlation of the volumes to be registered. Once calculated, the cross-correlation is normalised by the square root of the autocorrelation of each volume. Suited to volumes of the same modality, appealing properties of the metric include the

D

Box 1.

$$\frac{\partial S}{\partial p} = -\frac{\sum_{i=1}^{R}\left[\sum_{j=1}^{Q_i}\left(F\left(x_{ij}\right)\frac{\partial M\left(T\left(x_{ij},p\right)\right)}{\partial p}\right)-b\sum_{j=1}^{Q_i}\left(M\left(T\left(x_{ij},p\right)\right)\frac{\partial M\left(T\left(x_{ij},p\right)\right)}{\partial p}\right)\right]}{a}$$

$$= -\frac{\sum_{i=1}^{R}\left[\sum_{j=1}^{Q_i}\left(F\left(x_{ij}\right)\frac{\partial M\left(T\left(x_{ij},p\right)\right)}{\partial p}\right)-b\sum_{j=1}^{Q_i}\left(M\left(T\left(x_{ij},p\right)\right)\frac{\partial M\left(y\right)}{\partial y}\bigg|_{y=T\left(x_{ij},p\right)}\frac{\partial T\left(x_{ij},p\right)}{\partial p}\right)\right]}{a}$$

production of a search space containing sharp peaks and well defined troughs. The accurate alignment of volumes results in values near one being produced by the metric. Misalignment, in contrast, produces values of less than one. The distributed normalised correlation measure of similarity *S(F,M,T)* is defined as:

$$S\left(F,M,T\right) = -1\frac{\sum_{i=1}^{R}\left[\sum_{j=1}^{Q_i}\left(F\left(x_{ij}\right)-M\left(T\left(x_{ij}\right)\right)\right)\right]}{\sqrt{\sum_{i=1}^{R}\left[\sum_{j=1}^{Q_i}F\left(x_{ij}\right)^2\sum_{j=1}^{Q_i}M\left(T\left(x_{ij}\right)\right)^2\right]}} \tag{1}$$

where *F* and *M* are fixed and moving segment intensity functions respectively, *T* is a spatial transform, and x_{ij} is the j^{th} voxel of segment *i* from the fixed volume. *R* is the number of segment a volume is divided into and Q_i is the number of valid voxels mapped between segments identified by *i*. The derivative of the distributed similarity metric with respect to transform parameter *p* is computed as shown in Box 1.

In keeping with the equation shown in Box 1:

$$a = \sqrt{\sum_{i=1}^{R}\left[\sum_{j=1}^{Q_i}F^2\left(x_{ij}\right)\sum_{j=1}^{Q_i}M^2\left(T\left(x_{ij},p\right)\right)\right]}$$

and

$$b = \frac{\sum_{i=1}^{R}\left[\sum_{j=1}^{Q_i}F\left(x_{ij}\right)M\left(T\left(x_{ij},p\right)\right)\right]}{\sum_{i=1}^{R}\left[\sum_{j=1}^{Q_i}M^2\left(T\left(x_{ij},p\right)\right)\right]}$$

where $M(T(x_{ij},p))$ represents a discrete input which has been interpolated using a B-spline interpolation scheme. $\partial T(x_{ij},p)/\partial p$ is the transform Jacobian which is used to estimate variations in the mapped voxel coordinates with respect to transform parameter *p*. Importantly, when a voxel location that maps outside of the moving segment is encountered, its contribution to the local derivative is discarded.

Figure 3 shows results, plotted as time against number of Worker Agents, obtained whilst performing volume registration with normalised correlation as a similarity metric. As can be seen, average processing time of registration was reduced from 68 minutes to approximately 10 minutes when 10 Worker Agents were employed. The distributed algorithm was observed to converge after 54 iterations with transform parameters which matched those computed by the sequential algorithm.

FUTURE TRENDS

Deformable registration is based on the assumption that an evenly spaced mesh of control points can be placed over fixed and moving volumes. A transform is then used to estimate the displacement necessary to map between corresponding control point locations. In order for registration to be performed, each control point plus underlying intensities are transformed and compared with their counterpart until an acceptable level of similarity is achieved. Such algorithms have recently emerged in distributed processing environments and are achieved by assigning subsets of control points to individual processing nodes. Conveniently, the use of free-form deformations permits the independent movement of control points and removes the need for processor-to-processor communications (Loeckx, Maes, Vandermeulen, & Suetens, 2004). Crucially, the resulting implementations are designed to address the computational burden associated with the estimation of transform parameters for locally deformed volumes.

CONCLUSION

Registration is an important step in medical analysis tasks where information is extracted from a combination of sources. Traditionally, volume registration algorithms have been implemented using single processor architectures which are limited by memory and speed constraints. Such limitations have a negative impact in clinical applications where real-time processing allows physicians to monitor the progress of treatment while a patient is present. Importantly, the most successful registration algorithms employ intensity-based correlation as a measure of similarity (Maintz & Viergever, 1998; Zitova & Flusser, 2003). Although a variety of similarity metrics have been developed, in practice they represent a considerable computational burden during the alignment process. The main reason for this is the high cost associated with multiple evaluations of a complex transform and the interpolation of non-discrete intensity coordinates. The surveyed literature makes clear that concurrent similarity calculation can be achieved and provides better performance than non-parallel approaches (Christensen, 1998; Salomon et al., 2005; Warfield et al., 1998). The large speedups reported are, however, difficult to obtain and only specialised hardware is capable of maintaining such efficiencies when scaled. As a consequence, the applications developed to address the problem of slow volume registration speeds are restricted to high-cost specialised architectures found predominantly in the research environment.

While various approaches to distribution have been employed, it is fine-grained parallelism that achieves the best results (Hastings, Kurc, Langella, Catalyurek, Pan, & Saltz, 2003; Ourselin, Stefanescu, & Pennec, 2002; Rohlfing & Maurer, 2003). These methods are based on the low level decomposition of an algorithm within a tightly-coupled architecture. Understandably, such algorithms are difficult to implement and minimise computational expense by eliminating the exchange of data between processors. In this article, a novel coarse-grained approach to high performance intensity-based volume registration has been presented. The approach adopted has been shown to significantly improve processing speed, clearly outperforming sequential versions of the same algorithm. By employing an intensity-based algorithm, the complicated segmentation of features fundamental to landmark-based registration methods have been avoided. As a consequence, the algorithm distributed does not require user intervention. Unlike other high performance volume registration applications, the decoupling of algorithm components allows transforms of any type to be incorporated. This explicit separation also permits different similarity calculation and transform parameter optimisation strategies to be employed with only minor modifications to the existing framework.

REFERENCES

Christensen, G.E. (1998). MIMD vs. SIMD parallel processing: A case study in 3D medical image registration. *Parallel Computing, 24,* 1369-1383.

Hastings, S., Kurc, T., Langella, S., Catalyurek, U., Pan, T., & Saltz, J. (2003). Image processing for the grid: A toolkit for building grid-enabled image processing applications. In *Proceeding of the 3rd IEEE/ACM International Symposium on Cluster Computing and the Grid* (pp. 36-43).

Loeckx, D., Maes, F., Vandermeulen, D., & Suetens, P. (2004). Non-rigid image registration using free-form deformations with a local rigidity constraint. *Lecture Notes in Computer Science, 3216,* 639-646.

Maintz, J.B.A., & Viergever, A. (1998). A survey of medical image registration. *Medical Image Analysis, 2,* 1-36.

Nicolescu, C., & Jonker, P. (2000). Parallel low-level image processing on a distributed memory system. In *Proceedings of the 15th Workshop on Parallel and Distributed Processing* (pp. 226-233).

Nolle, L., Wong, K.C.P., & Hopgood, A.A. (2001). DARBS: A distributed blackboard system. *Research and Development in Intelligent Systems, 18,* 161-170.

Ourselin, S., Stefanescu, R., & Pennec, X. (2002). Robust registration of multi-modal images: Towards real-time clinical applications. In *Proceedings of the 5th International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 140-147).

Penney, G.P., Weese, J., Little, J.A., Desmedt, P., Hill, D.L.G., & Hawkes, D.J. (1998). A comparison of similarity measures for use in 2D-3D medical image registration. *IEEE Transactions on Medical Imaging, 17,* 586-595.

Pluim, J.P., Maintz, J.B.A., & Viergever, M.A. (2003). Mutual information-based registration of medical images: A survey. *IEEE Transactions on Medical Imaging, 22,* 986-1004.

Rohlfing, T., & Maurer, C.R. (2003). Non-rigid image registration in shared-memory multi-processor environments with application to brains, breasts, and bees. *IEEE Transactions on Information Technology in Biomedicine, 7,* 16-25.

Salomon, M., Heitz, F., Perrin, G.R., & Armspach, J.P. (2005). A massively parallel approach to deformable matching of 3-D medical images via stochastic differential equations. *Parallel Computing, 31*, 45-71.

Seinstra, F.J., Koelma, D., & Geusebroek, J.M. (2002). A software architecture for user transparent parallel image processing. *Parallel Computing, 28,* 967-993.

Stark, D.D., & Bradley, W.G. (1999). *Magnetic resonance imaging* (3rd ed.). Mosby Publishing, USA.

Tait, R.J., & Schaefer, G. (2008). *Similarity metrics for medical image registration*. This volume.

Tait, R.J., Schaefer, G., & Hopgood, A.A. (2006a). iDARBS—A distributed blackboard system for image processing. In *Proceedings of the 13th International Conference on Systems, Signals and Image Processing* (pp. 431-434).

Tait, R.J., Schaefer, G., & Hopgood, A.A. (2006b). Towards high performance image registration using intelligent agents. In *Proceedings of the 13th International Conference on Systems, Signals and Image Processing* (pp. 435-438).

Warfield, S.K., Jolesz, F., & Kikinis, R. (1998). A high performance approach to the registration of medical imaging data. *Parallel Computing, 24*(9), 1345-1368.

Yoo, T.S. (2004). *Insight into images: Principles and practices for segmentation, registration, and image analysis.* A.K. Peters Ltd, USA.

Zitova, B., & Flusser, J. (2003). Image registration methods: A survey. *Image and Vision Computing, 21*, 977-1000.

KEY TERMS

Blackboard Architecture: An artificial intelligence application based on and analogous to a group of experts seated in a room with a large blackboard working as a team to solve a common problem.

Coarse-Grained Parallelism: A term used to describe an algorithm that has been divided into high-level components, each of which can be hosted by a separate processor.

Distributed Agents: Software entities designed to execute as independent threads and on distributed processors, capable of acting autonomously in order to achieve a predefined task.

Fine-Grained Parallelism: A term used to describe an algorithm that has been divided into low-level components, each of which can be hosted by a separate processor.

Image Registration: The process whereby two images, differing by a spatial transformation, are brought into geometric alignment.

Parallel Processing: The concurrent execution of the same task, split into components, on multiple processors in order to achieve faster processing speeds.

Similarity Metric: A measure used to quantitatively judge how well a transformed sensed (moving) image fits a reference (fixed) image by comparing intensities.

Echocardiographic Image Sequence Compression Based On Spatial Active Appearance Model

Sándor Miklós Szilágyi
Hungarian Science University of Transylvania, Romania

László Szilágyi
Hungarian Science University of Transylvania, Romania

Zoltán Benyó
Budapest University of Technology and Economics, Hungary

INTRODUCTION

Echocardiography is a popular medical imaging modality due to its noninvasive and versatile behavior. There are no known side effects, and the measuring equipment is small and inexpensive relative to other options, such as MRI or CT. Reducing storage requirements and making data access user friendly are two important motivations for applying compression to ultrasound images, with the retention of diagnostic information being critical (Chiu, Vaisey, & Atkins, 2001).

A typical echocardiography image consists of a nonrectangular scanned area and a passive background, which may contain patient related text or limited graphics (e.g., a single channel ECG signal). The resulting spatial variation in image statistics presents a hard task to coding methods that use a single partition strategy. For example, many modern image compression algorithms, such as zero-tree coding (Shapiro, 1993) and set partitioning in hierarchical trees (SPIHT) (Said & Pearlman, 1996), are based on the wavelet transform, which partitions the input images into frequency bands whose size decreases logarithmically from high frequencies to low ones. This kind of decomposition strategy works well when the input images are statistically homogeneous, but not in the case of echocardiography image sequences.

Usually the medical applications do not tolerate much loss in fidelity, so the distortion free methods, such as context-based adaptive lossless image coding (CALIC) (Wu & Memon, 1997) have been recently adapted to "near-lossless" situations (Wu & Bao, 2000) with good results. Erickson, Manduca, Palisson, Persons, Earnest, and Savcenko (1998) have compared SPIHT and JPEG methods to compress magnetic resonance imaging (MRI) and ultrasound images. They concluded that wavelet-based methods are subjectively far superior to JPEG compressed at moderately high bit rates. Medical images are typically stored in databases, so it is possible for computers to extract patterns or semantic connections based on a large collection of annotated or classified images. Such automatically extracted patterns can improve the processing and classifying performance of the computers.

In the recent past, researchers in the image analysis community have successfully used statistical modeling techniques to segment, classify, annotate and compress images. Particularly, variations of hidden Markov models (HMMs) have been developed and successfully applied for image and video processing. The key issue in using such complex models is the estimation of system parameters, which is usually a computationally expensive task. In practice, often a tradeoff is accepted between estimation accuracy and running time of the parameter estimation method (Joshi, Li, & Wang, 2006).

Such a statistical information-based estimation highly depends on biological parameters. In our case, the most important task in efficient echocardiography image compression is the accurate detection of QRS complexes from the simultaneously measured ECG signal. Due to the quasiperiodic behavior of the ECG signal and echocardiography image sequences, the parameters of the patient model can be more precisely estimated.

Active appearance models (AAM), introduced by Cootes, Edwards, and Taylor (2001), are promising image segmentation tools that may provide solutions to most pending problems of echocardiography, as they rely on both shape and appearance (intensity and/or texture) information. Bosch et al. (2002) proposed a robust and time-continuous delineation of 2-D endocardial contours along a full cardiac cycle, using an extended AAM, trained on phase-normalized four-chamber sequences.

To understand the physiology and patho-physiology of the heart, not only the electrical activity and spatial distribution of its structures is important, but also their movement during normal and abnormal cardiac cycles. The ECG signal is measured simultaneously with echocardiography sequence recording in order to localize the investigated events (Szilágyi, Szilágyi, & Benyó, 2007b).

Several papers in the literature have already reported the usage of spatial AAM (Mitchell, Bosch, Lelieveldt, van der Geest, Reiber, & Sonka, 2002; Stegmann & Pedersen, 2005). The present work has the following contributions:

- We developed a heart reconstruction algorithm including time dependent wall boundaries in order to estimate the image variances that allow a better compression rate at a fixed image quality than conventional methods.
- Reported techniques classify ultrasound images only as belonging to systolic or diastolic interval. Our approach distinguishes normal and extra beats, and processes the corresponding images accordingly.

MATERIALS AND METHODS

Simultaneous echocardiography sequence recording and ECG signal measurement were carried out using a 2-D echocardiograph that produces 30 frames per second, and a 12-lead ECG monitoring system that samples at 500 Hz frequency and 12-bit resolution. Each image frame received a time stamp, which served for synchronization with ECG events. Two different series of measurements were recorded. The first series, which served for AAM training, consisted of 35 patients (12 of whom having extraventricular beats), 20 ultrasound sequences for each patient, of 10-15 seconds length each sequence, with previously established transducer placements. Based on these data, an a priori information database was created, which organized the ultrasound images grouped by corresponding ECG events.

The second series of measurements, which involved eight patients, consisted of two stages. In the first stage, the same measurements were performed as were performed in the first series, in order to provide patient-specific training data for the AAM. In the second stage, several measurements were performed using different placements and positions of the transducer. In this order, image sequences were recorded at eight parallel cross sections in horizontal and rotated (45° to the left and to the right) positions with a 1 cm inter-slice distance. We used 10 common axis planes that were placed at front, lateral and back side of the torso. For each patient a total number of $10 \times 8 \times 3 = 240$, at least 2-3 second long image sequences were created.

The duration of the recorded image sequences was restricted by the quasiperiodic behavior of the ECG signal. The spatial movement of the heart is constrained by the course of the depolarization-repolarization cycle (Szilágyi, Szilágyi, & Benyó, 2007a). For example, normal and ectopic beats imply different spatial heart movements. The studied ECG parameters were: shape of QRS beat, QT, and RR distances. These parameters characterize the nature of a QRS complex, and were determined as presented in Szilágyi et al. (2007a). ECG event clustering was accomplished using Hermite functions and self-organizing maps (Lagerholm, Peterson, Braccini, Edenbrandt, & Sörnmo, 2000). Two main event clusters were created: normal and ventricular extra beats. This latter group, because of the patient specific manifestation of ventricular extras, had to be dealt with separately patient by patient. QRS beats not belonging to any cluster were excluded from further processing, together with their corresponding ultrasound sub-sequences. A further condition for normal QRS complexes to be included was having RR distance between 700-800 ms and QT distance between 350-400 ms. A detailed presentation of ECG processing is presented in section (a) in Figure 1 (Szilágyi et al. 2007a).

The time-varying evolution of the cardiac volume is determined by the interconnection of electrical and mechanical phenomena. In a whole cardiac cycle there are two extremity values. The maximal volume can be coupled with the starting moment of ventricular contraction. The moment of minimal volume shortly precedes the termination of ventricular contraction, but is much more difficult to identify, due to the dead time of a normal cardiac cell. This delay is caused by the property of a regular cardiac cell, whose electric response precedes with 60-80 ms the mechanical contraction (Winslow, Hinch, & Greenstein, 2005). The combination of the electrical and mechanical properties of the heart and the usage of knowledge-base allowed us to create a performance evaluation module that determines the most probable wall position, as shown in section (b) of Figure 1.

Figure 1, section (b) presents an overview of the image processing and volumetric reconstruction procedure. The first algorithmic step is noise elimination. Speckle noise represents a major difficulty to most ultrasound imaging applications (Evans & Nixon, 1996). In our case, the suppression of such phenomena was accomplished using the well-known motion adaptive spatial technique presented in Szilágyi et al. (2007b). Due to the measuring technique of traditional echocardiography, the obtained images are distorted. In order to become suitable input for 3-D processing, they need to go through a normalization transform. Every recorded ultrasound slice is represented by a plane whose spatial alignment depends on the position and rotational angle of the transducer. The normalization process also takes into consideration the distance of each image pixel from the transducer.

The training data of the AAM is constructed based on the spatial position of each echocardiography slice recorded in the first measurement series. By averaging these spatial distributions, a mean base 4-D heart shape model is obtained, which will be the starting point of the AAM (Bosch et. al., 2002; Lelieveldt, van der Geest, Mitchell, Bosch, Sonka & Reiber, 2002). Landmark points are determined using the technique proposed by Mitchell et al. (2002). The sparse character of the obtained spatial description model doesn't allow the landmark points have 3-D texture information. That is why we restricted the texture to 2-D.

Subjects have their own specific, time dependent inner structure, which cannot be approximated properly

Figure 1. (a) Schematic representation of the ECG signal filtering, processing and compression, (b) Schematic representation of the data recording and analyzing procedure. All echocardiography and ECG data go through the same processing module. The AAM is constructed from the measurements of series 1, and fine tuned afterward using the patient specific data resulting from series 2 stage 1. Stage 2 data serve for the detailed cardiac volumetric analysis. Reconstructed 3-D objects are finally aligned using an iterative LMS-based algorithm.

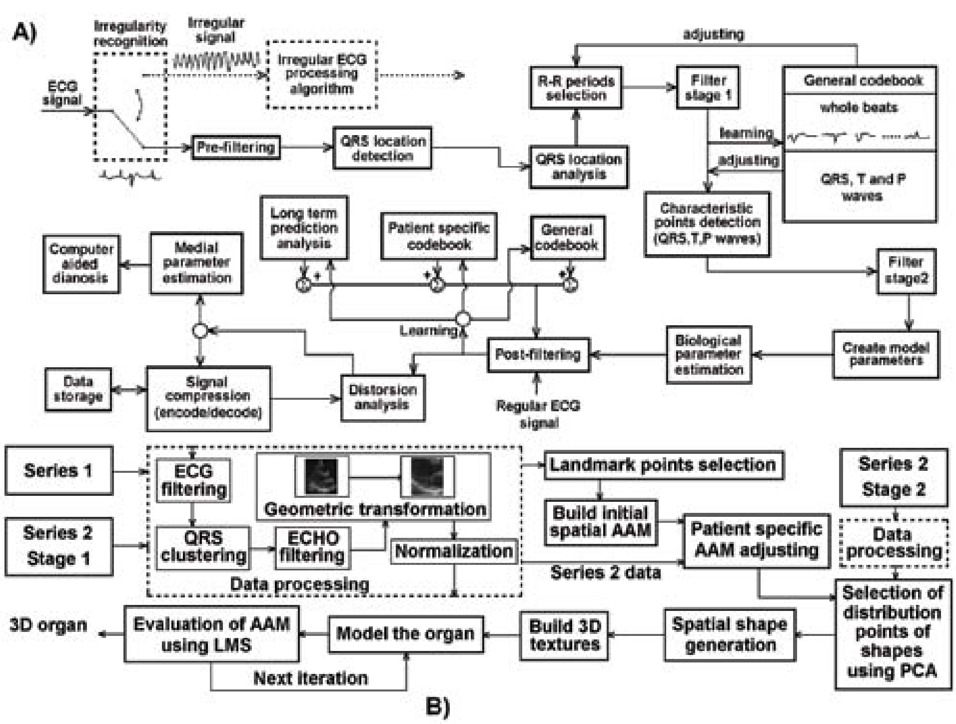
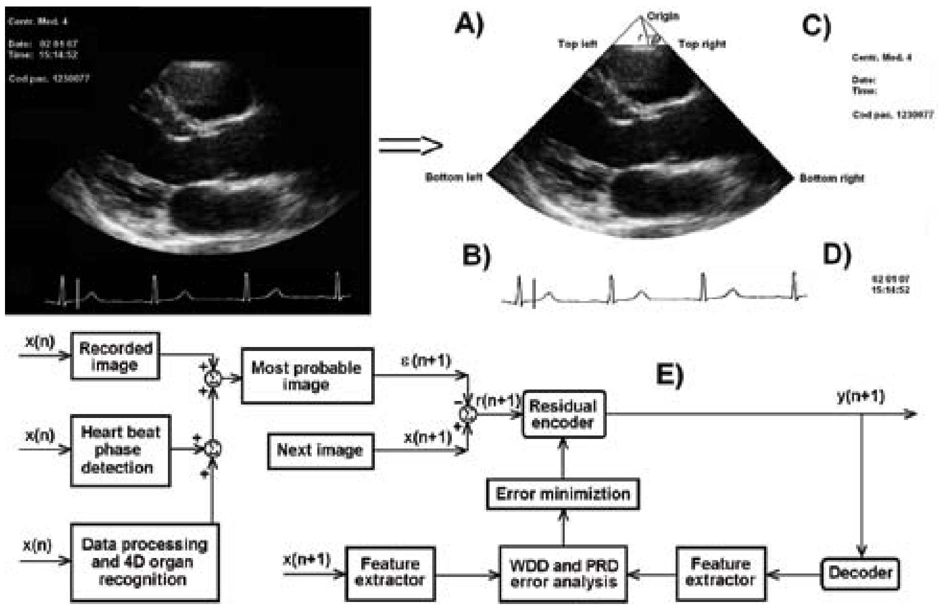


Figure 2. Separation of an echocardiographic scene into: (a) nonrectangular echocardiographic image area, (b) one lead of ECG signal, (c) static text and graphic labels, and (d) variable text. In section (a) is presented the origin point, the distance r from origin and the angle θ. The origin's coordinate is deductible from positions of bottom left-right and top left-right points. (e) Schematic representation of residual data construction and distortion analysis. The residual signal calculation process uses the recorded image, heart beat phase information and AAM-based 4D organ shape to determine the most probable image in the next measuring moment. As the y(n+1) output signal is obtained, a WDD and PRD-based distorsion analysis is realized, where the image features are determined from AAM.



from a population of few dozens of individuals. In order to make further adjustments to the AAM, the base structure was adjusted to the patient using the measurements made in the second series, first stage. The landmark points determined on the images recorded during the second stage of the second measurement series allow us to create a 3-D distribution point model, which was established according to (Mitchell, 2002). Having the distribution points established, the AAM will be enabled to adjust itself to a diversity of biological factors like the phase of ECG and breathing. A detailed description of the manifestation of these phenomena and the model adaptation is given in (Stegmann & Pedersen, 2005). Our algorithm acts similarly, but it treats the cardiac cycle differently: not only systolic and diastolic phases are distinguished, but also a QRS complex clustering is performed to give different treatment to normal and ventricular cardiac cycles.

Spatial texture maps are determined via averaging (Stegmann & Pedersen, 2005). The visual aspect of the heart and its environment, because of their mutual motion, is changing in time. AAM models only include information on the texture situated within the model. The time dependent representation of the ultrasound slices obtained from the large stack of sequences enabled us to accurately determine the 4-D structure of the heart (Mitchell, 2002). The iterative algorithm of the AAM demands the comparison of measured and expected shapes. The AAM was adjusted using a quadratic cost function, until the desired accuracy was obtained.

Due to its adaptive behavior, the compression method is able to handle patient-dependent data, and has a capability to efficiently separate the measured artifacts from the useful signal.

The proposed signal compression algorithm can be divided into the following steps:

- Intelligent image sequence analysis and filtering (that involves the automatic recognition of echocardiography image, ECG signal, and various changing and constant labels and letters that appear on the recorded image sequence);

- Background selection (the constantly dark region);
- ECG signal processing (see section (a) of Figure 1);
- Segmentation of echocardiographic image (see section (b) of Figure 1);
- Calculation of the heart's 4D shape (3D+time) using AAM;
- Estimation of probable image;
- Residual signal estimation, entropy coding (encode and decode), and back-estimation.

The compression evaluation was implemented in two ways, using percentage root mean square difference (PRD) and weighted diagnostic distortion (WDD) (Zigel, Cohen, & Katz, 2000). WDD measures the relative preservation of the diagnostic information (such as location, duration, intensity, shapes, edges) in the reconstructed image. These diagnostic features were determined by physicians. The distortion estimation and signal (ultrasound image and all auxiliary data) coding process is presented in Figure 2e.

As sections (a) through (d) of Figure 2 present, it is possible to separate logically the various image regions and in function of its behavior apply a separate compression scheme.

RESULTS

Section (a) of Figure 3 presents two series of ultrasound slices indicating the contour of the left ventricle of the second patient, detected during a ventricular contraction. The two rows of slices show two different angle views, having a 60° angle difference. The four slices in each row represent subsequent images of the sequence, showing the approximately 100 ms duration of the ventricular contraction. Section (b) of Figure 3 shows two different reconstructed 3-D shapes of the left ventricle.

Figure 4 presents the decoded and feature-base corrected echocardiographic images at various compression rates. The variation of RMSE level against inverted compression rate is shown in section (e) of Figure 4. The RMSE graph was created for a normal QRS beat shape and average RR and QT distances. We can observe a pronounced performance advantage of the method for normal beats due to its higher incidence.

DISCUSSION

The recognition of the relation between echocardiography images and simultaneously recorded ECG signal is a key element in efficient image sequence compression (see Table 1). However, various events like aspiration and expiration may influence the measured data. During a whole cardiac cycle, the shape and volume of the left ventricle changes considerably. It is difficult to determine the performance of the reconstruction method for the sporadically occurred ventricular extra-systolic beats. Even for patients that produce at least five extra beats with similar shapes in each minute the reconstruction performance remains well below the normal QRS cluster's accuracy, due to the sparse distribution of the processable slices.

Figure 3. Results of shape reconstruction: (a) Time varying 2 D contour of the left ventricle, (b) reconstructed 3-D structure of the left ventricle

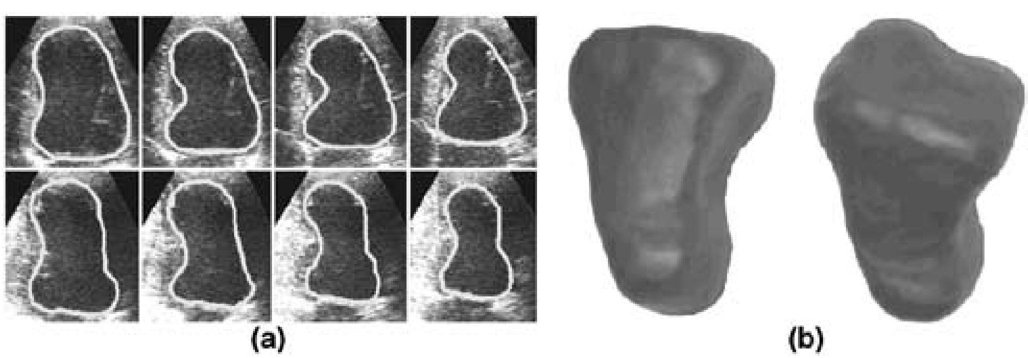
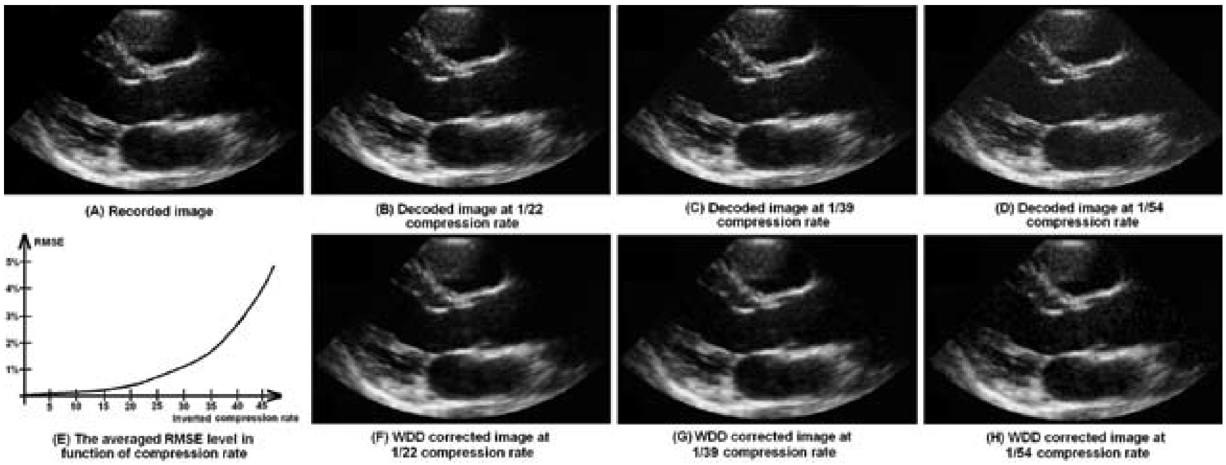


Figure 4. Results of image reconstruction: (a) Original image, (b) decoded image at 1/22 compression rate, (c) decoded image at 1/39 compression rate, (d) decoded image at 1/54 compression rate, (e) representation of root mean square error (RMSE) in function of inverted compression rate, (f) feature-base corrected image, b, (g) feature-base corrected image, c, (h) feature-base corrected image, d



From image 4 we can observe that even a well-working WDD correction method cannot handle a compression rate better than 1/60 without a serious image distortion that can lead to wrong medical diagnosis. Table 1 demonstrates the higher perfomance of the image compression method, that exist due to the advanced QRS beat analysis and spatial AAM-based organ reconstruction. Such an analysis can lead to a much better *estimated image* quality that reduces the amplitude of the residual signal.

CONCLUSION

The investigation of simultaneously recorded ECG and echocardiography images enables us to study the relations between the electrical and mechanical phenomena concerning the heart. The compression method presented in this article performs well in case of normal and almost as well for ventricular beats. This kind of approach to the problem may result in deeper understanding of electrical and mechanical properties of the heart that enable a much more efficient compression than the results obtained from algorithms using less a priori information.

ACKNOWLEDGMENT

This research was supported in part by the Sapientia Institute for Research Programmes, Domus Hungarica Scientiarium et Artium, and the Communitas Foundation.

Table 1. The obtained inverted compression rates obtained for normal beats using: (a) Wavelet-based compression, (b) Image comparison-based estimation, (c) QRS, long-term prediction (LTP) ECG and echocardiographic image compression, (d) AAM-based compression

RMSE\ Method type	Wavelet -based	Image comparison	QRS and LTP Image comparison	AAM-based
1%	10.2	14.2	17.4	22.7
2%	15.4	21.2	26.6	36.3
3%	17.3	24.0	29.6	40.2
4%	18.8	26.1	31.8	44.1

REFERENCES

Bosch, J. G., Mitchell, S. C., Lelieveldt, B. P. F., Nijland, F., Kamp, O., Sonka, M., et al. (2002). Automatic segmentation of echocardiographic sequences by active appearance motion models. *IEEE Trans. Med. Imag., 21*, 1374-1383.

Chiu, E., Vaisey, J., & Atkins, M. S. (2001). Wavelet-based space-frequency compression of ultrasound images. *IEEE Trans. Inf. Tech. Biomed. Eng., 5*, 300-310.

Cootes, T. F., Edwards, G. J., & Taylor, C. J. (2001). Active appearance models. *IEEE Trans. Patt. Anal. Mach. Intell, 23*, 681-685.

Erickson, B. J., Manduca, A., Palisson, P., Persons, K. R., Earnest, D., & Savcenko, V. (1998). Wavelet compression of medical images. *Radiology, 206*, 599-607.

Evans, A. N., & Nixon, M. S. (1996). Biased motion-adaptive temporal filtering for speckle reduction in echocardiography. *IEEE Trans. Med. Imag., 15,* 39-50.

Joshi, D., Li, J., & Wang, J. Z. (2006). A computationally efficient approach to the estimation of two- and three-dimensional Hidden Markov Models. *IEEE Trans. Image Proc., 15*, 1871-1886.

Lagerholm, M., Peterson, C., Braccini, G., Edenbrandt, L., & Sörnmo, L. (2000). Clustering ECG complexes using Hermite functions and self-organizing maps. *IEEE Trans. Biomed. Eng. 47*, 838-848.

Lelieveldt, B. P. F., van der Geest, R. J., Mitchell, S. C., Bosch, J. G., Sonka, M., & Reiber, J. H. C. (2002). 3-D active appearance models: Fully automatic detection of endoand epicardial contours in short-axis cardiac MR data. *Proc. Int. Soc. Magn. Res. Med. (ISMRM), 2*, 1668.

Mitchell, S. C., Bosch, J. G., Lelieveldt, B. P. F., van der Geest, R. J., Reiber, J. H. C., & Sonka, M. (2002). 3-D active appearance models: Segmentation of cardiac MR and ultrasound images. *IEEE Trans. Med. Imag. , 21*, 1167-1178.

Said, A., & Pearlman, W. A. (1996). A new, fast, and efficient image codec based on set partitioning in hierarchical trees. *IEEE Trans. Circ. Syst. Video Techn., 6*, 243-250.

Shapiro, J. (1993). Embedded image coding using zerotrees of wavelet coefficients. *IEEE Trans. Sign. Proc., 41*, 3445-3462.

Stegmann, M., & Pedersen, D. (2005). Bi-temporal 3D active appearance models with applications to unsupervised ejection fraction estimation. *Progr. Biomed. Opt. Imag. Proc. SPIE, 5746*, 336-350.

Szilágyi, S. M., Szilágyi, L., & Benyó Z. (2007a). Support vector machine-based ECG compression. *Advances in Soft Computing, Springer (IFSA 2007), 41*, 737-745.

Szilágyi, S. M., Szilágyi, L., & Benyó, Z. (2007b). Volumetric analysis of the heart using echocardiography. *Lect. Notes Comp. Sci. (FIMH 2007), 4466,* 81-90.

Winslow, R. L., Hinch, R., & Greenstein, J. L. (2005). Mechanisms and models of cardiac excitation-contraction coupling. *Lect. Notes Math., 1867*, 97-131.

Wu, X., & Bao, P. (2000). L-infinity constrained high-fidelity image compression via adaptive context modeling. *IEEE Trans. Image Proc., 9*, 536-542.

Wu, X., & Memon, N. (1997). Context-based adaptive lossless image coding. *IEEE Trans. Comm., 45*, 437-444.

Zigel, Y., Cohen, A., & Katz, A. (2000). ECG signal compression using analysis by synthesis coding. *IEEE Trans. Biomed. Eng., 47*, 1308-1316.

KEY TERMS

Active Appearance Model (AAM): A computer vision algorithm for matching a statistical model of object shape and appearance to a new image.

ECG Analysis: A collection of ECG signal processing methods aiming at the determination of medical parameters.

Echocardiogram: Image of the heart created with ultrasound imaging technology.

Hidden Markov Model (HMM): Astatistical model in which the system being modeled is assumed to be a Markov process with unknown parameters, and the challenge is to determine the hidden parameters from the observable parameters.

Image Compression: Compression of the data contained by an images. The objective is to reduce the image's redundancy, which assures an efficient storage or transmission of the recorded data.

QRS Complex: A structure on the ECG signal that corresponds to the depolarization of the ventricles.

Weighted Diagnostic Distortion: A weighted measurement of signal distortion in term of changes of medical parameters.

Electrical Bioimpedance Cerebral Monitoring

Fernando Seoane
University College of Borås, Sweden

Kaj Lindecrantz
University College of Borås, Sweden

INTRODUCTION

Electrical Bioimpedance (EBI) is now a mature technology in medicine, with applications in clinical investigations, physiological research, and medical diagnosis (Schwan, 1999). The first monitoring application of bioimpedance techniques, *impedance cardiography*, date back to 1940. Since then, bioimpedance measurements have been used in several medical applications, from lung function monitoring and body composition, to skin cancer detection. A complete historical review is available in Malmivuo and Plonsey (1995). A medical imaging modality based on bioimpedance, *Electrical Impedance Tomography* (EIT) has also been developed (Bayford, 2006).

EBI has been used to study the effect in the brain of spreading depression, seizure activity, asphyxia and cardiac arrest since 1950s and 1960s (Ochs & Van Harreveld, 1956), but the most important activities in electrical cerebral bioimpedance research has been during the last 20 years (Holder, 1987; Holder & Gardner-Medwin, 1988). Examples of areas of study are *brain ischemia*, *spreading depression*, *epilepsy*, *brain function monitoring*, *perinatal asphyxia*, *monitoring of blood flow, and stroke*.

BACKGROUND

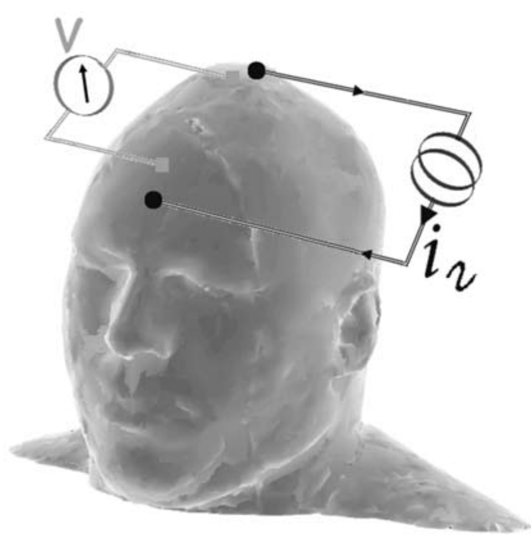
The basic functional units of an EBI measurement system for cerebral monitoring are the following: an electric current generator, a voltage meter, the surface electrodes for current injection and voltage pick up as well as the connecting electrical leads. The injected current causes a voltage drop in the tissue that is sensed and the measured bioimpedance is calculated from the resulting quotient from voltage over current, know as Ohm's Law. See Figure 1 and Equation 1.

$$Z(\omega) = \frac{V(\omega)}{I(\omega)} \quad (1)$$

An important feature of the application of EBI for cerebral monitoring is that it is applicable in some of the situations where brain is particularly at risk as well as for long-term monitoring situations where available imaging techniques: *MRI, CT-scan* are not suitable (*e.g., during an ongoing cardiopulmonary by-pass operation, in intensive care, for acute stroke assessment in ambulances*).

Other features of bioimpedance technology are that it is harmless for the patient, portable and very affordable in comparison with other monitoring techniques already in used. These specific features place EBI as the technology of choice to fill the need for brain monitoring in the medical scenarios mentioned above and several others.

Figure 1. Functional diagram of a measurement system for electrical bioimpedance cerebral monitoring



Electrical Properties of Biomaterials and Bioimpedance

Biological material, tissue, and cells have electrical properties (conductivity *σ* and permittivity *ε*) that allow electrical current to flow in the presence of an electric field (Cole, 1968). These electrical properties depend of the constitutive elements and structure of tissue; therefore, changes in structure or biochemical composition modify the electrical properties, *σ* and *ε,* of the tissue, and consequently the electrical impedance changes.

Every type of tissue and body fluids in the human body exhibits specific values of conductivity and permittivity that are frequency dependent. Therefore, each tissue can be characterized by its particular electrical impedance spectrum, and measurements of electrical impedance can be used to differentiate between tissues or to assess the state of the tissue. The most complete compilation of the dielectric properties of biological tissues is found in (Gabriel, 1996).

Electrical Impedance Tomography

EIT exploits the fact that the electrical properties differ between tissue types to create an image representing the conductivity distribution of a volume conductor. Despite its name, EIT does not reconstruct and object slice by slice, because electrical current cannot be confine in to a plane; instead, the current will flow through the whole volume conductor following the gradient of the electrical field.

As several conductivity distributions may provide the same voltage boundary detected by the sensing electrodes, reconstruction of the impedance image requires that some assumptions are made, and also that a model is used for fitting the voltage boundary data.

There are two methods for EIT imaging: absolute and difference imaging. The first one, also known as static method, obtains a conductivity image from a set of impedance measurements. The difference imaging method uses a set of two measurements taken at two different times to create a conductivity image of the differences. This method is used for dynamic studies, for monitoring changes in the tissue.

EIT imaging exhibits a poor spatial resolution as compared to other imaging techniques but the time resolution, in the order of microseconds, is unique for EIT. Affordable, portable, and noninvasive are other exclusive features to EIT within brain imaging monitoring. For a deep understanding of EIT, see a recent review (Bayford, 2006).

Electrical Impedance Spectroscopy Analysis

One straight-forward effect of the frequency dependency of electrical properties of tissue is the effect of the cellular membrane. Figure 2 shows the influence of the frequency on the current path lines. The capacitive effect of the membrane contributes to the electrical properties of tissue and depends of many factors: the number of cells, the size of the cells, the thickness of the cell membranes, type of cells, and so on.

Because the plasma membranes of the tissue cells act as a capacitive element, most of a direct current (DC) in biological tissue, flows through the extracellular space (*e.g., interstitial fluid, plasma, and so on*). Hence, the impedance at DC is mainly determined by the conductivity of the extracellular fluid, the available surface to the electrical field for the charges to flow through, and the length of the propagation path. See Figure 2 and Equation 2. At higher frequencies the capacitive effect shunts the electrical current, allowing the electrical current to propagate also via the intracellular space and the tissue conductivity can be modelled as in Equation 3.

$$\boldsymbol{g}_{DC} = \sigma_e \frac{2(1-\boldsymbol{f})}{(2+\boldsymbol{f})} \qquad (2)$$

$$\boldsymbol{g} = \sigma_e \frac{2(1-\boldsymbol{f})\,\sigma_e + (1+2\boldsymbol{f})\left(\dfrac{\sigma_i\,(\sigma_m + i\omega \boldsymbol{C}_m)a}{\sigma_i + (\sigma_m + i\omega \boldsymbol{C}_m)a}\right)}{(2+\boldsymbol{f})\,\sigma_e + (1-\boldsymbol{f})\left(\dfrac{\sigma_i\,(\sigma_m + i\omega \boldsymbol{C}_m)a}{\sigma_i + (\sigma_m + i\omega \boldsymbol{C}_m)a}\right)} \qquad (3)$$

Equations legend: $\boldsymbol{g}$ is the complex conductivity of tissue, $\boldsymbol{\sigma}_e$ is conductivity of the extracellular fluid, S×m^{-1}, $\boldsymbol{\sigma}_i$ is conductivity of the intracellular fluid S×m^{-1}, $\boldsymbol{\sigma}_m$ is the membrane conductivity, S×m^{-1}, $\boldsymbol{c}_m$ is the surface membrane capacity, Farads×m^{-2}, ω is the angular frequency, radians×s^{-1}, $\boldsymbol{i}$ is the imaginary number √-1, $\boldsymbol{a}$ is the cell radius and $\boldsymbol{f}$ is the volume fraction of concentration of cells.

N.B. the expression for $\boldsymbol{g}_{DC}$ corresponds to the general expression for $\boldsymbol{g}$ when ω = 0 and $\boldsymbol{\sigma}_m$ is approximate to 0.

Notice in Equation 3 that the conductance of tissue depends not only on the dielectric properties of the constituents but its morphology and frequency. For this reason, each tissue has specific electrical properties and consequently presents a specific impedance spectrum. This fact allows the differentiation between tissues and assessment of health state by spectrum analysis of the tissue impedance. This is the principle behind several impedance-based medical applications (*e.g.*, *skin cancer screening* (Aberg, Nicander, Hansson, Geladi, Holmgren, & Ollmar, 2004), and *breast cancer assessment* (Assenheimer & Laver-Moskovitz, 2001)).

Potential Applications

There are several areas of neurology that may benefit from the application of EBI technology as a means to obtain useful indicators about the undergoing activity of brain threatening mechanism.

Brain Monitoring during Cardiac Surgery

The high vulnerability of brain tissue to hypoxia (Acker & Acker, 2004) together with the ability of impedance measurements to detect hypoxic cell swelling (Seoane, Lindecrantz, Olsson, Kjellmer, Flisberg, & Bågenholm, 2005) place EBI technology as a candidate method for early detection of hypoxic brain damage during cardiac surgery.

Perinatal Asphyxia

It has been proven (Seoane et al., 2005) that the cell swelling associated with hypoxic-ischemic injury can be sensed and measured via EBI.

Brain Function and Epilepsy

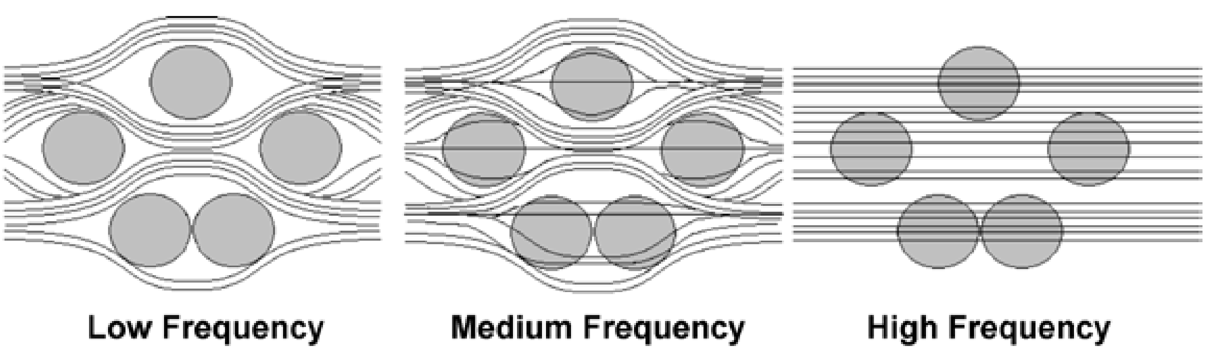
The electrical properties of the cell membranes are nonlinear (Hodgkin, 1947) and during neural depolarization the opening of the voltage-sensitive ion channels modify the membrane impedance (Holder, 1992).

The principle behind detection of epilepsy using EBI measurements relies on the fact that the impedance of an epileptic region will be appreciably different of the rest of the brain because of the large electrical activity associate to the epileptic seizure, accompanied by local cell swelling and ischemia (Olsson & Broberg, 2006).

Stroke

There are two kinds of stroke, ischemic stroke, and haemorrhagic stroke. Both result in alterations in the composition and the structure of the affected area. In the case of ischemic stroke the blood flow to an area of the brain is interrupted and consequently there is a lack of blood and oxygen that results in ischemic cell swelling. In the case of haemorrhagic stroke, the rupture of blood vessels allows the leaking of blood from the cerebrovascular system to accumulate in the intercellular space causing a haematoma. These two different sequences of pathophysiological events alter the electrical properties of the affected area and the alteration can be detected by means of noninvasive electrical impedance measurements (Liu & Dong, 2006).

Figure 2. The effect of the cell membrane capacitance in the current path lines



Cerebral Blood Flow Monitoring

Autoregulation of cerebral blood blow is a protective mechanism that stabilises cerebral perfusion during changes in blood pressure, often induced as a consequence of intracranial hypertension or cerebral ischemia. The use of intracranial measurements of EBI can detect cerebral blood blow autoregulation (Bodo, Pearce, Baranyi, & Armonda, 2005), suggesting that rheoencephalography could be developed into a non-invasive method for early detection of brain injury.

CURRENT RESEARCH ISSUES IN ELECTRICAL BIOIMPEDANCE CEREBRAL MONITORING

Electrical bioimpedance technology provides useful features for cerebral continuous monitoring and encouraging results have been seen in various clinical applications. However, there are still many issues to be addressed before the technology is developed into a clinical tool for cerebral monitoring.

Biophysical Understanding

Generally, there is a lack of understanding about the relationship between the underlying pathophysiological mechanisms associated to a particular insult (*e.g., cell swelling, intracranial haemorrhage, and so on*), and the influence on the electrical properties of tissue. The knowledge about impedance dynamics and the injury or adaptation mechanisms is still coarse. The ultimate effect of the damaging insult on the electrical properties of the tissue is relatively well-known, and the fundamental hypotheses are supported by experimental results. However, the effect on the electrical properties of the tissue constituents, of the biochemical and histological processes active during the corresponding adaptation or injury mechanism is, in most cases, completely unknown.

Impedance Data for Tissue Spectra Characterization

The specific electrical properties of each biological tissue confer an impedance spectrum virtually unique to each particular tissue. Available biological dielectric data is compiled and available in (Gabriel, 1996). As the origin of the data is mostly excised animal samples and human corps, the values of these data are not necessarily representative for in vivo tissue.

A proper spectral characterization of healthy tissue will empower the identification and classification applications based in EBI spectroscopy. It is essential to have spectral information of healthy tissue, as well as enough spectral data to fully characterize the impedance spectra of the injured tissue.

Current Density Distribution in the Head

There are several factors that determine the current density distribution within the head. These factors influence the current density distribution directly or indirectly. The ultimate direct factor is the electrical impedivity distribution in the head. The frequency and the placement of electrodes are factors that indirectly affect the current density distribution.

Effect of the Skull

When current is injected into the head, the current density is much smaller in deep intracranial region than in the superficial layers, skin, and scalp (Malmivuo, Suihko, & Eskola, 1997; Seoane, Lu, Persson, & Lindecrantz, 2007). Recent simulations with a realistic model have shown that the effect of the skull is less shielding than earlier believed (Seoane et al., 2007), and it has been shown that the current density distribution depends largely on the electrical properties of each specific tissue. A smaller current density in the intracranial structures is explained, not only by the shielding layer of low conductive bone. The conductive extracranial layer of muscle tissue, together with the large area available for the current to flow within the brain, are also responsible for the small intracranial current density.

Volume under Study, Placement of Electrode and Sensitivity Maps

Electrode placement significantly influences in the current density distribution, as seen in Figure 3, thus a particular arrangement of electrodes will be better for the study of a certain region. Not only the position of the pair of injecting electrodes, but also the position of the pair of sensing electrodes is important. The arrangement of the injecting and sensing electrodes determines the sensitivity map; therefore, different placement of the pairs of electrodes gives different sensitivity maps. These maps may contain positive and negative as well as null regions (*e.g., if the pair of electrodes were overlapping each other, the sensitivity map will be only positive*). For this reason, some electrode arrangements are more suitable for study of certain phenomena than others. An adjacent placement of electrodes will give information mainly from the cortex, desirable for the study of certain pathologies (*e.g., epilepsy*), while the same electrode arrangement will provide very little information from the centre of the brain, less favourable for detection of haemorrhage in the brain ventricles for instance.

In EIT, the placement of electrodes is determined by the selected measurement strategy: opposite method, cross method, neighbouring method, and adaptive method. The corresponding sensitivity maps associated with each of them have been studied in Kauppinen, Hyttinen, and Malmivuo (2006).

Spatial Resolution in EIT

The main limitation of EIT is the poor spatial resolution. The accuracy of the image has improved as new processing methods have been introduced, but it is still poor compared to traditional imaging methods used in clinical practice like MRI, CT, Ultrasound, and so on.

Number of Channels

A progressive increase in the number of measurement channels of the EIT systems, from 16 electrodes used back in 1987, up to with 128 electrodes (Gang, Lim, George, Ybarra, Joines, & Liu, 2006), has contributed to improve the quality and the resolution of the impedance images (Mengxing, Wei, Wheeler, McCormick, & Xiuzhen, 2002), as expected.

Reconstruction Algorithms

The reconstruction algorithm contains several simplifications and assumptions that worsen the accuracy of the reconstructed image, thus the selection of a specific algorithm and its components are critical for the spatial resolution of an EIT image.

The complexity of the models used for the forward problem has been increased from simple heterogeneous 2D-spherical models to 3D-layered concentric sphere, and even realistic anatomic models based in MRI-images (Liston, Bayford, Tidswell, & Holder, 2002). The increase in the complexity of the models has been accompanied with the corresponding improvement in resolution and quality of the image (Bagshaw & Liston, 2003; Liston et al., 2002).

The work done in reconstruction algorithms for EIT during the last decade and the generated scientific literature is immense, and part of it is contained in the review work done by Lionheart (2004).

FUTURE TRENDS

Researchers and physicians have realized of the potential of EBI technology for monitoring of the brain; consequently, the proliferation of bioimpedance techniques targeting the dynamics of basic pathophysiological mechanism is natural and expected.

A better biophysical understanding of the dynamics of EBI during the aetiologies of interest along with a good knowledge regarding the current density distribution and the impedance sensitivity maps in the brain will most probably lead to the development of aetiology-specific operation modes in EBI measurement systems.

The Promising Magnetic Induction EIT

Magnetic Induction Electrical Impedance Tomography, MI-EIT, is a modality of EIT. In MI-EIT, the current is not directly applied into the tissue through electrodes—it is magnetically induced by coils (Al-Zeibak & Saunders, 1993). This way, it is possible to reduce the effect of the low conductivity of the skull bone and eliminate certain limitations related to the electrodes (Scharfetter, Casanas, & Rosell, 2003).

MI-EIT has been under study and continuous development for the past years (Al-Zeibak & Saunders, 1993; Rosell, Casanas, & Scharfetter, 2001; Scharfetter et al., 2003), and recently, its application in cerebral monitoring has been intensified.

CONCLUSION

Changes in the electrical properties of brain tissue do reflect certain physiological activities in the brain. These activities may originate from normal processes (*e.g., the membrane depolarization during an evoked potential*), or it might be the result of brain damage (*e.g., ischemic oedema after stroke*), or it might be the injury mechanism itself (*e.g., hypoxic cell swelling*). Therefore, the use of electrical bioimpedance measurements of the brain can play an important role supporting early diagnosis of several brain-related conditions.

The development of electrical bioimpedance technology has been intense and continuous, especially during the last decade, but it has not been enough to reach the necessary status to be applied in clinical practice as brain monitoring tool yet. However, through experimental research results, hardware developments, and simulation studies, there may be a breakthrough soon.

REFERENCES

Aberg, P., Nicander, I., Hansson, J., Geladi, P., Holmgren, U., & Ollmar, S. (2004). Skin cancer identification using multifrequency electrical impedance - a potential screening tool. *IEEE Transactions on Biomedical Engineering, 51*(12), 2097–2102.

Acker, T., & Acker, H. (2004). Cellular oxygen sensing need in CNS function: Physiological and pathological implications. *Journal of Experimental Biology, 207*(18), 3171–3188.

Al-Zeibak, S., & Saunders, N. H. (1993). A feasibility study of in vivo electromagnetic imaging. *Physics in Medicine and Biology, 38*(1), 151–160.

Assenheimer, M., Laver-Moskovitz, O., et al. (2001). The T-SCAN™ technology: Electrical impedance as a diagnostic tool for breast cancer detection. *Physiological Measurement, 22*(1), 1–8.

Bagshaw, A. P., Liston, A. D., et al. (2003). Electrical impedance tomography of human brain function using reconstruction algorithms based on the finite element method. *NeuroImage, 20*(2), 752–764.

Bayford, R. H. (2006). Bioimpedance tomography (electrical impedance tomography). *Annu Rev Biomed Eng, 8*, 63–91.

Bodo, M., Pearce, F. J., Baranyi, L., & Armonda, R. A. (2005). Changes in the intracranial rheoencephalogram at lower limit of cerebral blood flow autoregulation. *Physiological Measurement, 26*(2), 1–17.

Cole, K. S. (1968). *Membranes, ions, and impulses. A chapter of classical biophysics.* (Vol. 1). Berkley: University of California Press.

Gabriel, C. (1996). *Compilation of the dielectric properties of body tissues at RF and microwave frequencies*. Brooks, TX: Air Force Material Command, Brooks Air Force Base.

Gang, Y., Lim, K. H., George, R., Ybarra, G., Joines, W. T., & Liu, Q. H. (2006). *A 3D EIT system for breast cancer imaging.* Arlington, VA.

Hodgkin, A. L. (1947). The membrane resistance of a nonmedullated nerve fibre. *J Physiol, 106*(3), 305–318.

Holder, D. S. (1987). Feasibility of developing a method of imaging neuronal activity in the human brain: A theoretical review. *Med Biol Eng Comput*, *25*(1), 2–11.

Holder, D. S. (1992). Impedance changes during the compound nerve action potential: Implications for impedance imaging of neuronal depolarisation in the brain. *Medical; Biological Engineering; Computing, 30*(2), 140–146.

Holder, D. S., & Gardner-Medwin, A. R. (1988). Some possible neurological applications of applied potential tomography. *Clinical Physics and Physiological Measurement*, *4A*, 111–119.

Kauppinen, P., Hyttinen, J., & Malmivuo, J. (2006). Sensitivity distribution visualizations of impedance tomography measurement strategies. *International Journal of Bioelectromagnetism, 8*(1), 9.

Lionheart, W. R. B. (2004). EIT reconstruction algorithms: Pitfalls, challenges and recent developments. *Physiological Measurement, 25*(1), 125–142.

Liston, A. D., Bayford, R. H., Tidswell, A. T., & Holder, D. S. (2002). A multi-shell algorithm to reconstruct EIT images of brain function. *Physiological Measurement, 23*(1), 105–119.

Liu, L. X., Dong, W., et al. (2006). A new method of noninvasive brain-edema monitoring in stroke: Cerebral electrical impedance measurement. *Neurol Res, 28*(1), 31–37.

Malmivuo, J., & Plonsey, R. (1995). *Bioelectromagnetism - principles and applications of bioelectric and biomagnetic fields*. New York: Oxford University Press.

Malmivuo, J., Suihko, V., & Eskola, H. (1997). Sensitivity distributions of EEG and MEG measurements. *Biomedical Engineering, IEEE Transactions on, 44*(3), 196–208.

Mengxing, T., Wei, W., Wheeler, J., McCormick, M., & Xiuzhen, D. (2002). The number of electrodes and basis functions in EIT image reconstruction. *Physiological Measurement, 23*(1), 129–140.

Ochs, S., & Van Harreveld, A. (1956). Cerebral impedance changes after circulatory arrest. *Am J Physiol, 187*(1), 180–192.

Olsson, T., Broberg, M., et al. (2006). Cell swelling, seizures and spreading depression: An impedance study. *Neuroscience, 140*(2), 505–515.

Rosell, J., Casanas, R., & Scharfetter, H. (2001). Sensitivity maps and system requirements for magnetic induction tomography using a planar gradiometer. *Physiological Measurement, 22*(1), 121–130.

Scharfetter, H., Casanas, R., & Rosell, J. (2003). Biological tissue characterization by magnetic induction spectroscopy (MIS): Requirements and limitations. *IEEE Transactions on Biomedical Engineering, 50*(7), 870–880.

Schwan, H. P. (1999). The practical success of impedance techniques from an historical perspective. *Ann N Y Acad Sci, 873* 1–12.

Seoane, F., Lindecrantz, K., Olsson, T., Kjellmer, I., Flisberg, A., & Bågenholm, R. (2005). Spectroscopy study of the dynamics of the transencephalic electrical impedance in the perinatal brain during hypoxia. *Physiological Measurement, 26*(5), 849–863.

Seoane, F., Lu, M., Persson, M., & Lindecrantz, K. (2007). *Electrical bioimpedance cerebral monitoring. A study of the current density distribution and impedance sensitivity maps on a 3D realistic head model*. Paper presented at the The 3rd International IEEE EMBS Conference on Neural Engineering.

KEY TERMS

Bioimpedance Tomography: Also known as Electrical Impedance Tomography, it is a medical imaging technique in which an image of the dielectric properties of biological tissue is inferred from surface electrical measurements.

Brain Stroke: A cerebrovascular accident that occurs as a consequence of brain ischemia or cerebral haemorrhage; ischemic stroke and haemorrahagic stroke respectively.

Electrical Bioimpedance: The physical magnitude that indicates the total impediment that a biomaterial offers to the flow of free electrical charges and the orientation of bounded electrical charges towards an existing electrical field.

Electrical Conductivity: A dielectric property that indicates the ability of a material to allow the flow of electrical charges.

Epilepsy: A neurological disorder in which abnormal electrical activity in the brain causes seizures.

Hypoxia: Reduction of oxygen supply to tissue below physiological level.

Ischemia: A low blood flow state leading to hypoxia in the tissue.

Seizure: Abnormal electrical activity in the brain tissue, which is usually accompanied by motor activity and/or sensory phenomena. Its origin can be epileptic or nonepileptic.

Electrical Impedance Spectroscopy as a Powerful Analytical Tool for Monitoring Microbiological Growth on Medical Implants

Louise M. Oliver
University of Ulster, UK

E.T. McAdams
University of Ulster, UK

P.S.M. Dunlop
University of Ulster, UK

J. A. Byrne
University of Ulster, UK

I.S. Blair
University of Ulster, UK

M. Boyle
Royal College of Surgeons, Ireland

K.G. McGuigan
Royal College of Surgeons, Ireland

INTRODUCTION

Hospital-acquired infections (HAI) are defined as infections that are neither present nor incubating when a patient enters the hospital (Bourn, 2000). Their effects vary from discomfort to prolonged or permanent disability and they may contribute directly or substantially to a patient's death. HAI's are estimated to cost the National Health Service (NHS) in England £1 billion annually (Bourn, 2000) with as many as 5,000 patients dying as a result of acquiring such an infection (Anon, 2001). Not all hospital-acquired infections are preventable but Infection Control Teams believe that they could be reduced by at least 15%, with yearly savings of £150 million (Anon, 2001). Central intravascular catheters have been found to be a common source of infection. Catheters can become infected via a number of different routes with the infection proliferating in multiple areas along the catheter surface. It has been reported that over 40% of the identified micro-organisms causing hospital-acquired infection were *Staphylococci*, an organism that is typically found on the natural skin flora (Bourn, 2000).

If an infection associated with the device is suspected, clinicians can either treat the infection with antibiotics or replace the catheter. Both of these methods increase patient discomfort, place an additional workload upon hospital staff, drain hospital resources, and increase the patient's length of stay within the hospital. It would be hugely beneficial to patients and clinicians if the onset of subclinical biofilm formation on medical implants could be detected, and thus treated, at an early stage.

The idea of using impedance to monitor bacterial growth is not a new one; indeed impedance microbiology was first introduced over 100 years ago (Stewart, 1878). Traditional impedance microbiology involves the application of a sinusoidal voltage or current typically across two-electrodes and the resulting impedance (or conductance) response is monitored at only one-frequency during the growth of the micro-organism under investigation. Although the area has shown promise, advances have been limited. The majority of published works have focused on changes in the bulk electrolyte with few studies examining the electrode-electrolyte interface and only a few recent studies have examined the development of sensors for monitoring bacterial growth (Gomez, Bashir, & Bhunia, 2002; Gomez et al., 2001; Oliver et al., 2006; Yang, Li, Griffis, & Johnson, 2004).

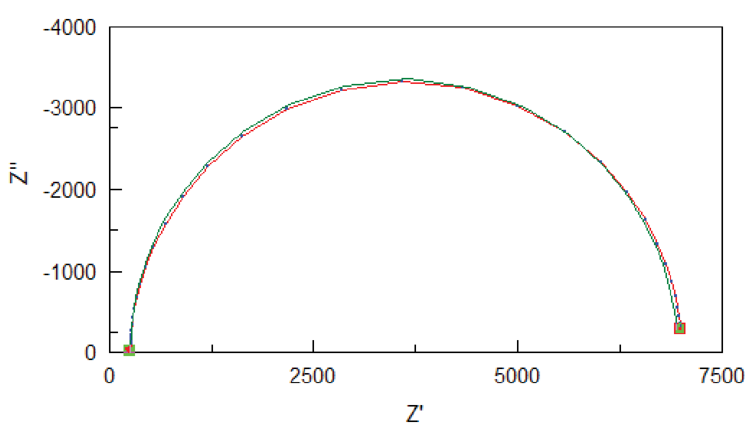
An alternative to monofrequency measurements is Frequency Response Analysis using Electrical Impedance Spectroscopy (EIS) and equivalent circuit modelling. When used to study electrochemical systems, EIS can give accurate, kinetic, and mechanistic information using a variety of techniques and output formats. For example, consider a small alternating voltage of known angular velocity, (ω) and small amplitude (v) applied to the system under investigation. The current response (i) and phase difference (φ) of the resultant waveform is measured for a range of applied frequencies (ω). The relationship between the applied voltage and resultant current waveform is the impedance (Z) of the system. During measurements, the amplitude of the applied perturbation waveform is carefully controlled to ensure minimal disturbance to the test system. EIS is therefore considered to be a nondestructive and safe method for studying biological systems. Established AC and electrochemical theory can be used to characterise the impedance of the electrochemical cell and to represent it in terms of an equivalent circuit model.

BACKGROUND: IMPEDANCE MICROBIOLOGY

The need to monitor bacterial growth effectively is apparent in the areas of food science, water quality, and medicine. However, there have been relatively limited advances in improving the means of detecting microbial growth. Conventional detection methods require long sample turnaround times, have low reproducibility, produce high error rates, and are costly in terms of labour (Noble, 1994). Automating microbiological detection would be advantageous as it would increase productivity and relieve human beings from having to perform mundane, menial, and repetitive tasks (Trotman, 1978). Impedance microbiology has been presented as a mechanistic alternative to monitoring the growth of bacteria and has become an accepted method of monitoring growth of micro-organisms in the brewing industry. Although this acceptance has been a promising advance, the monofrequency method used to date limits the potential of the technique. In the microbiology field, there have been relatively few studies which have explored the multifrequency impedance analysis method. It is proposed that a multifrequency approach will not only facilitate the monitoring of microbial growth rates but will also provide a means of monitoring microbial-environment interactions.

Electrical impedance spectroscopy differs from traditional impedance microbiology measurements in that it measures the impedance response over a wide range of frequencies and the data can be modelled using an equivalent circuit which can be interpreted in terms of the electrochemical reactions occurring within the system. The combination of a multifrequency approach, improved electrode design, and the use of meaningful equivalent circuit modelling enables the simultaneous examination of the electrode-electrolyte interface and

Figure 1. A typical complex impedance plot showing one partial semi-circular arc



the bulk properties of the electrolyte. Data is typically presented on a complex impedance plot (Z real component vs. Z imaginary component) and, generally, over the relatively limited frequency ranges available with the commercial devices (Hz to MHz), forms one or two partial semicircular arcs (Figure 1). Over an extensive frequency range several such arcs should be observed. The Bode Plot presents both the log |Z| and the phase angle (φ) against log frequency (ω) (Figure 2). The most appropriate circuit model is then applied to the data. It is important that data is viewed in a number of different plot formats so that the appropriateness of the circuit model can be established. In the case where the Randles equivalent circuit (Figure 3). is the model of choice, the bulk resistance of the solution is examined through analysis of the series resistance. The properties of the interface are examined by monitoring changes in the parallel resistance and capacitance (or constant phase element). Solution resistance dominates at high frequencies and the electrode-electrolyte interface at low frequencies, provided that data is collected over the appropriate frequency range.

THE BULK PROPERTIES OF THE MICROBIAL SOLUTION

The processes occurring in the bulk of the solution are very different from the processes that occur at the electrode surface. In solutions there is a two-way flow of ions, that is, positive ions are attracted to the negative cathode and negative ions are attracted to the positive anode. The movement of ions in an electrolyte is altered by the application of an electric field. The solution resistance measures the opposition to current flow travelling through the solution by the migration of ions and is represented in equivalent circuit modelling by the series resistance (R_s). The mobility of the ions directly affects the bulk resistance measurement.

Growth media consists of strong and weak electrolytes and includes proteins, carbohydrates and minerals. When the electrolyte concentration of the growth medium changes as a result of bacterial metabolism, larger molecules are broken down into smaller ions (Cady, 1978; Ur & Brown, 1975). This eases the flow of current through the solution/media and results in an increase in conductivity or a decrease in the bulk

Figure 2. Bode Plot for an electrochemical cell, showing the modulus of the impedance and the phase angle against the log of frequency

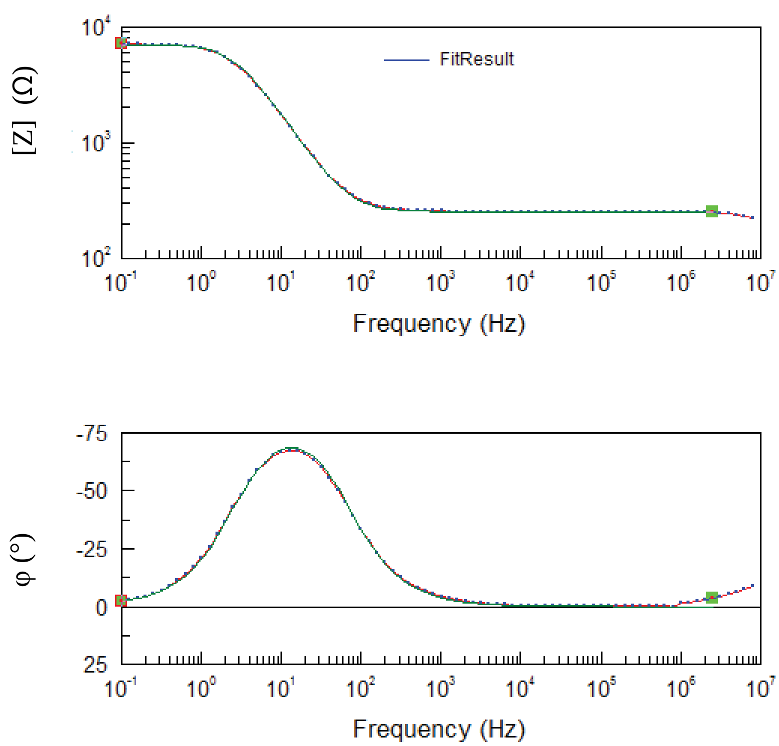
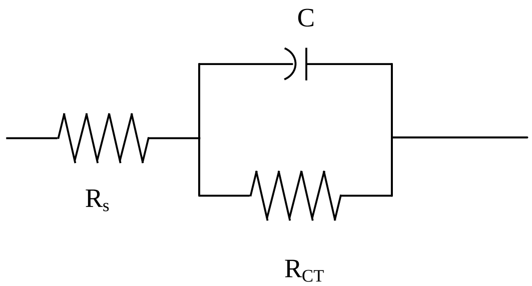


Figure 3. Equivalent circuit model sometimes referred to as the Randles circuit model



resistance. Given the large and varied amount of components of the growth medium, it is not possible to calculate the number of ions either completely ionised or dissociated during microbial growth. Owens (1985) had difficulty trying to accurately predict the change in conductivity as a function of carbon source, growth conditions, and pH buffers. However, it was his belief that conductivity could be maximised by considering separately the factors that both increase and decrease the conductance of a medium. The significance of this was that culture media could be formulated that maximised the observed changes in conductivity or bulk resistance. The choice of media is therefore considered of paramount importance when examining changes in the resistance of the bulk solution (R_s).

THE ELECTRODE-ELECTROLYTE INTERFACE IN MICROBIOLOGY

The environment at the electrode interface is very different from the growth medium/bulk solution. At the electrode-electrolyte interface, there is transference of charge from the electrons in the metal electrode to the ions in the solution. At the electrode-solution interface, an electric double layer is formed, which contains a nonuniform distribution of charges. Helmholtz (1853), as cited by Macdonald (2005), compared the double layer to a molecular parallel plate capacitor, where one plate represents the charges in the metal and the other plate represents the ions (at a minimum distance) in the solution. The Helmholtz model is only valid for high concentration electrolyte solutions. In more dilute solutions the transition will not be so abrupt and the thickness of the double layer increases. Helmholtz's model predicts that changes in capacitance are due to dielectric permittivity or the distance between the plates. Alternative models have been developed that do not agree with Helmholz's theories. Two studies (Chapman, 1913; Gouy, 1909) proposed that an electrode submerged in a liquid is surrounded by an ionic cloud rather than considering the interface between an electrode and a liquid as rigidly fixed charge as Helmoltz had proposed. In this case the counterion atmosphere will be more like the ionic atmosphere around an individual ion which is termed the diffuse layer. Gouy and Chapman's model takes into account thermal motion which influences the equilibrium of the distribution of counterions. The ions around the electrode diffused exponentially into the liquid medium as distance increased from the electrode (Macdonald, 2005). The limitations of the Gouy-Chapman model have been identified (Bockris & Reddy, 1970) where it was established that the Gouy-Chapman model estimates capacitances with reasonable accuracy provided solutions have very low numbers of ions. Stern (1924) amalgamated the Helmholtz and Gouy-Chapman models by suggesting that the diffuse layer is divided into an inner layer (Stern, 1924) and an outer layer (Chapman, 1913; Gouy, 1909). There have been a number of modifications and extensions to the original Helmholtz model, thus indicating the complexities of the electrode-electrolyte interface in the field of electrochemistry.

In equivalent circuit modelling, the double layer of charge at the electrode interface is often termed the double layer capacitance C_{DL}. The DC current that manages to cross the electrode-electrolyte interface experiences a charge transfer resistance, R_{CT} where the expression for R_{CT} (Equation 1) is derived from the Butler-Volmer equation used in electrochemistry;

$$R_{CT} = \frac{RT}{nF}\frac{1}{i_0} \qquad (1)$$

Where R is the gas constant, T is the temperature (K), F is the Faraday constant, n is the number of electrons involved in the electrode reactions, and i_0 is the exchange current density.

For systems that exhibit ideal capacitive behaviour at the electrode-solution interface, the estimation of the double layer capacitance is relatively simple using techniques such as impedance spectroscopy. In practice, most systems do not exhibit ideal behaviour and show noncapacitive behaviour at the interface as a result of surface roughness on the electrode and diffusion/adsorption at the electrode interface (Bard & Faulkner, 2001; Macdonald, 2005). "Distributed" elements such as the Warburg element and the constant phase element (CPE) prove useful when describing these non-ideal processes. The Warburg element describes diffusion related events and the CPE accounts for the non-ideal capacitive properties of the system. Diffusion related events are rarely seen in the impedance microbiology area and for this reason will not be described here. Many theories have been proposed in electrochemistry to describe the electrode-solution interface; few theories have emerged to describe the electrode-bacterial interface.

ELECTRICAL IMPEDANCE SPECTROSCOPY OF MICROBIAL BIOFILM FORMATION

EIS was used to monitor the growth of *Staphylococcus* epidermidis RP62A over a period of 24 hours. Circuit modelling provided information on the growth mechanisms of *S. epidermidis.* An electrode system designed using a flexible implantable substrate (Polyimide) is tested and presented as a possible system for monitoring the growth of *S. epidermidis* on medical implants. All experiments were thermostatically controlled using an incubator at 37°C ± 0.1°C and sterile brain heart infusion (BHI) media (Oxoid). Culture dishes containing 2 ml sterile BHI media were inoculated with 200 μL of an overnight culture of *S. epidermidis*. Electrodes were immersed in the sterile media two hours prior to inoculation. Electrodes were fabricated by deposition of gold on polyimide (PI). 125 µm Du Pont Kapton™ was plasma cleaned using helium plasma (60 W for 2 minutes). A 20 nm titanium interlayer was sputter-deposited on the polyimide, followed by a 200 nm layer of gold. A series of photolithographic steps were used to form the electrode pattern shown in Figure 4. The electrode array was designed so that it could be used for two and three-electrode techniques although only the results obtained using a two-electrode technique are presented here. The two-electrode technique measured the impedance response between electrodes B and C (electrode A was not used for measurements). Electrodes B and C are identical in size (1 mm^2) and shape.

Before impedance measurements the electrodes were cleaned with Piranha solution ($H_2SO_4 : H_2O_2$ 3:1) then rinsed thoroughly with sterile distilled water and dried in N_2 atmosphere. An oscillating AC potential of 20 mV was applied to the system (0 V DC-bias) over the frequency range 10 Hz to 1 MHz using a Solartron 1260 frequency response analyzer combined with a Solartron 1287 electrochemical interface. Impedance measurements were taken every 30 minutes for a period of 24 hours following inoculation of BHI media with *S. epidermidis*. A control cell was setup containing sterile BHI media without bacteria and measurements were taken at the same intervals for comparison. All measurements and analysis was performed using the Zplot and Zview software provided by Solartron Inc.

The charge transfer resistance is considered to be in series with the solution resistance R_s and in parallel with the double layer capacitance. Commercial software packages are available (e.g., Gamry, Scribner and Solartron) providing equivalent circuits models such as Figure 5 to model the experimental data. A constant phase element (CPE) was used instead of a pure capacitance to fit the experimental data because biological responses produce a depressed semi-circular arc, where the intercept angle at high frequencies is not 90° as expected for a pure capacitance. The use of CPE for biological systems has produced much better fits with experimental data than a pure capacitance as the electrode/solution interface exhibits non-capacitive properties (McAdams & Jossinet, 1996). The impedance of a CPE is defined in equation 2 where *K* is a measure of the magnitude of Z_{CPE} and has units of Ω s$^{-\alpha}$, j is the imaginary unit ($j = \sqrt{-1}$), ω is the angular frequency ($\omega = 2\pi f$) and α is a constant which has a value between 0 and 1. Any complex plot where $\alpha<1$ is more accurately modelled by a CPE. There is no

Figure 4. Miniature EIS probes made by gold deposition onto polyimide

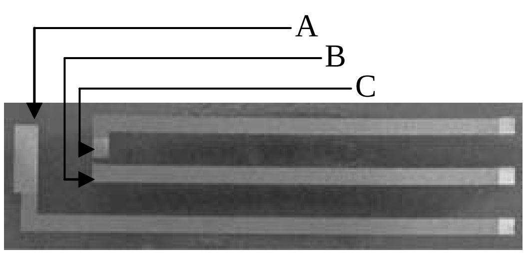


Figure 5. Equivalent circuit model, showing the bulk resistance R_S, the charge transfer resistance R_{CT} and a constant phase element CPE

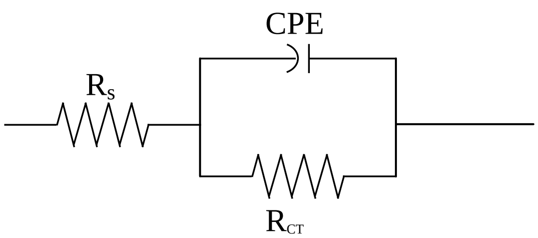
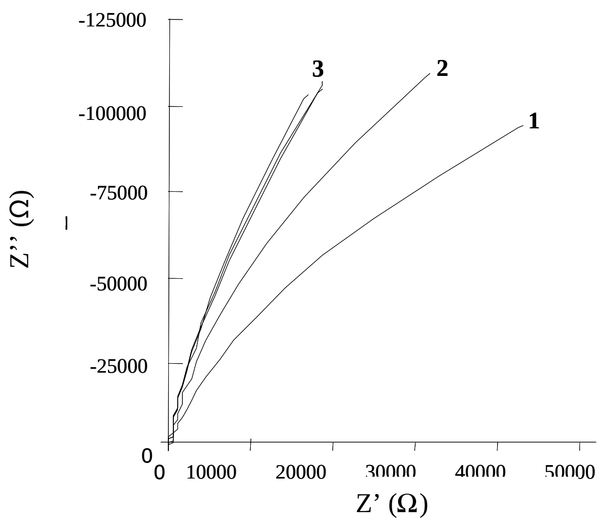


Figure 6. Complex impedance plots (resistance Z' vs. reactance Z") for sterile media (1), inoculation (2) and during biofilm growth (3)



possibility of estimating the double layer capacitance C_{DL} from K values (Zoltowski, 1998).

$$Z_{CPE} = \frac{1}{K(j\omega)^{\alpha}} \qquad (2)$$

It is important to remember that the CPE exhibits no capacitive behaviour and the assumption that the double layer capacitance C_{DL} = K is incorrect and results in inaccuracies in all other modelled parameters (McAdams, 1996; MacDonald, 2005; Zoltowski, 1998).

RESULTS

Staphylococcus epidermidis RP62A adhesion to the gold electrode surface was visually confirmed following impedance measurements by staining the biofilm with crystal violet. This procedure stains the biofilm cells deep purple. Figure 6 shows plots for blank media, inoculation point (t = 0 h) and during microbial biofilm formation. The shape of the data plots are different for pre and post inoculation. The major differences are observed between the blank and culture plots at the lower frequency end of the spectrum, where charge transfer resistance (R_{CT}) dominates. Figures 7-9 show the modelled data for all complex plots for the 24 hour measurement period. The bulk resistance (R_s) of the media decreased over the 24 hour measurement period as shown in Figure 7 at t = 0 h R_s = 264 Ω and at t = 24 h R_s = 170 Ω, a decrease of 94 Ω. This is explained by the theory proposed by Cady (1978) which states that bacterial metabolism causes the breakdown of large organic molecules in the media into smaller more conducive ions which gives an increase in media conductivity over time. The decrease in bulk resistance was not as marked in the blank media, however R_s did show a decrease with time, at t=0 R_s=187Ω and at t=24 R_s=175 Ω, a decrease of 12 Ω.

Values for K did not change markedly over time and remained relatively constant at 0.15 μΩ$s^{-\alpha}$ for the inoculated solution (Figure 8). Values of K for the blank media were higher (0.2 μΩ $s^{-\alpha}$) compared to that of the inoculated media (0.15 μΩ $s^{-\alpha}$) at all time points, which shows that the presence of the bacteria decreased the impedance of the CPE at the interface. R_{CT} was observed to increase dramatically in the first few hours of biofilm formation (Figure 9). A magnitude of 0.6 MΩ was recorded at inoculation which doubled after 2 hours to 1.2 MΩ and continued to increase for the duration of the experiment. There was a small initial decrease in R_{CT} for the blank media in the first 30 minutes and thereafter the value remained constant at ca. 0.1 MΩ for the duration of the experiment. It is important to note that the complex impedance plots for this set of experiments (Figure 6) are not complete semicircular arcs and therefore values produced by the commercial software for R_{CT} are an approximation and subject to error. Comparing the data for R_{CT} during biofilm growth with R_{CT} values for blank media there is a marked difference in their pattern and magnitude (Figure 9). Further investigation at lower frequencies is required

E

Figure 7. R_s vs. Time for sterile BHI media (□) and during biofilm formation (■)

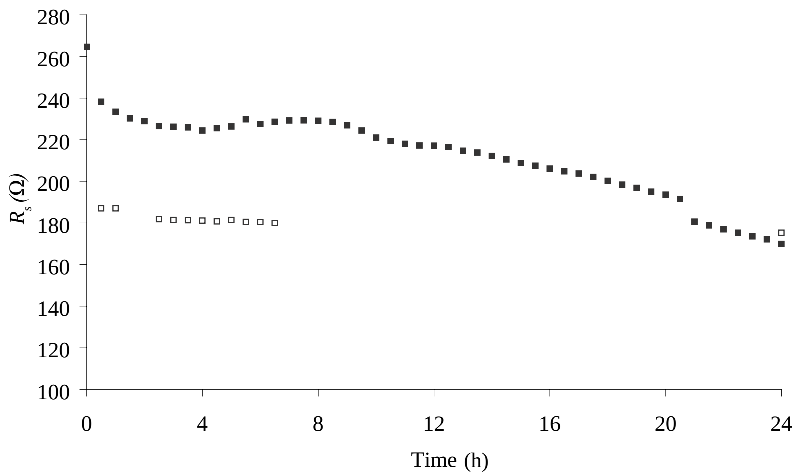


Figure 8. K vs. Time for sterile BHI media (□) and during biofilm formation (■)

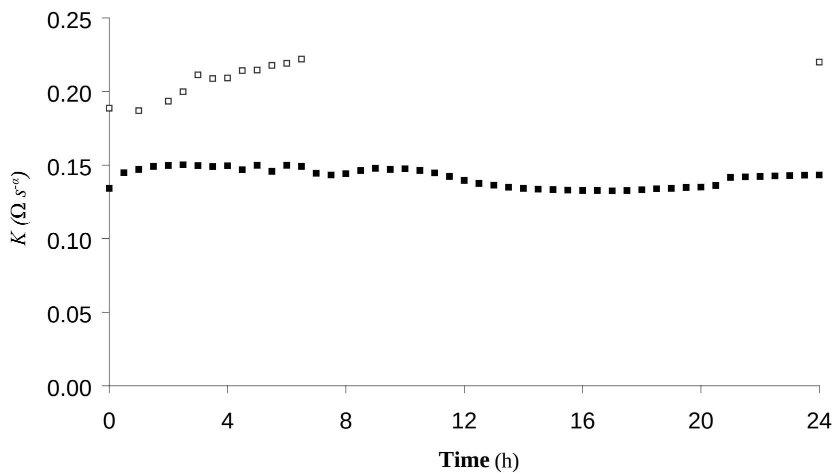
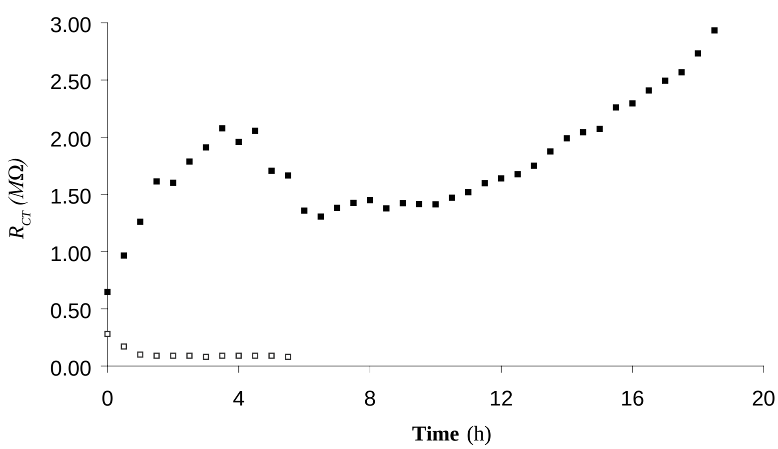


Figure 9. R_{CT} vs. Time for sterile BHI media (□) and during biofilm formation (■)



to determine if a change in R_{CT} could be used to detect initial biofilm adhesion and recruitment.

CONCLUSION

Microbial biofilms are defined as a matrix-enclosed bacterial population adhered to a surface or each other. An organic matrix protects these sessile organisms (Dominguez-Benetton & Castaneda, 2005; Costerton, 1995; Marshall, 1971; Poulsen, 1999). Biofilms are characterized in part by the production of a network of highly hydrated extracellular polysaccharides (EPS). The EPS matrix has a number of functions, one of which includes facilitating initial attachment of bacteria to the surface. The EPS matrix has been reported to change the surface charge and free energy on the surface to be colonized and may account for changes observed in R_{CT} (Poulsen, 1999). Changes in R_{CT} can be explained by microbial adhesion where adhesion to the surface is usually characterized by two phases (Marshall, 1971): (i) The reversible phase which occurs when the bacteria are very near to the surface but not in direct contact and includes electrostatic, hydrophobic interactions, and Van der waals forces, (ii) the irreversible phase where binding to the surface is complete and includes dipole-dipole interaction and covalent bonding. It is possible that changes in R_{CT} in the first few hours of microbial growth (Figure 9) relate to the reversible stage of the biofilm formation with the continuing increase in R_{CT} relating to the second phase of biofilm formation. An increase in R_{CT} was not observed in the sterile media. However, partial arcs in complex plots do not allow for a good approximation of R_{CT}, and future work should include an investigation of lower frequencies (<10 Hz). Changes in the double layer are often related to biofouling and biocorrosion, of metal electrodes (Dominguez-Benetton and Castaneda 2005). In this set of experiments, the gold surface of the electrode did not become corroded. Within the field of biocorrosion, biofilm formation is monitored sometimes over a period of thirty days and in future experiments impedance monitoring will take place over longer periods of time. The observed decrease in the magnitude of R_s over time follows Cady's theory where the breakdown of metabolites in the media due to bacterial respiration and growth causes an increase in the net conductance of the media and consequently a decrease in R_s. Future sensor design should take into account that the biofilm does not adhere well to the substrate, preferring the gold electrodes; this biological parameter may be exploited to improve the sensitivity of the impedance technique. However, one must also be careful not to promote biofilm recruitment on implantable devices. The electrode-medium interface warrants further investigation to determine the most effective parameter for monitoring biofilm formation and bacterial growth. This study confirms that electrical impedance spectroscopy is a possible method for the in-situ or possible *in vivo* monitoring of biofilm recruitment and/or bacterial growth.

FUTURE TRENDS

The use of EIS and equivalent circuit modelling to monitor microbial growth has shown great promise for routine microbiology and also in medical applications. In routine detection it offers the ability to monitor microbial interactions in the bulk solution and at the electrode interface with potentially much faster response times than traditional methods and with a greater degree of automation. More work will be required using different strains of micro-organism to confirm these exciting early findings. In order for EIS to be routinely used by microbiologists, there would need to be a simplification of the technology and further studies could exploit developments in micro-electronics to produce more portable impedance monitoring systems.

REFERENCES

Anon. (2001). *Surveillance of hospital-acquired bacteraemia in English hospitals 1997-2001*. London: Public Health Laboratory Service.

Bard, A.J., & Faulkner, L.R. (2001). *Electrochemical methods: Fundamentals and applications* (2nd ed.). Wiley.

Bockris, J., & Reddy, O.M. (1970). *Modern electrochemistry.* New York: Plenum Press.

Bourn, J. (2000). *The managment and control of hospital acquired infection in acute NHS trusts in England*. London: National Audit Office.

Cady, P. (1978). Impedimetric screening for bacteriuria. *Journal of Clinical Microbiology, 7,* 273-278.

Chapman, D.L. (1913). A contribution to the theory of electro-capillarity. *Physiological Magazine, 25*, 475-481.

Costerton, J.W. (1995). Microbial biofilms. *Annual Review of Microbiology, 49*, 711-745.

Dominguez-Benetton, X., & Castaneda, H. (2005). SRB-biofilm growth and influence in corrosion monitoring by electrochemical impedance spectroscopy. *Corrosion 2005, 5486,* 1-10.

Gomez, R., Bashir, R., & Bhunia, A.K. (2002). Microscale electronic detection of bacterial metabolism. *Sensors and Actuators B, 86,* 198-208.

Gomez, R., Bashir, R., Sarikaya, A., Ladisch, M.R., Sturgis, J., Robinson, J.P. et al. (2001). Microfluidic biochip for impedance spectroscopy of biological species. *Biomedical Microdevices, 3*(3), 201-209.

Gouy, G. (1909). Sur la constitution de la charge electrique a la surface d'un electrolyte. *Comptes rendes hebdomadaires des Seances de l'Academie des Sciences, 149,* 654-657.

Helmholtz, H. (1853). Ueber einige Gesetze der Vertheilung elekrischer Strome in korperlichen Leitern mit Anwendung auf die theirisch-elektrischen Versuche. *Annals of Physics, 89*, 211-233.

McAdams, E.T., & Jossinet, J. (1996). Problems in equivalent circuit modelling of the electrical properties of biological tissues. *Bioelectrochemistry and Bioenergetics, 40*, 147-152.

Macdonald, J.R. (2005). *Impedance spectroscopy emphasizing solid materials and systems* (2nd ed.). New York: Wiley.

Marshall, K.C. (1971). Mechanisms of the initialevents in the sorption marine bacteria to surfaces. *Journal of General Microbiology, 68,* 337-348.

Noble, P.A. (1994). *Theoretical and practical aspects of impedance microbiology*. Ph.D. Thesis: University of Saskatchewan.

Oliver, L.M., Dunlop, P.S.M., Byrne, J.A., Blair, I.S. M. Boyle, K.G. McGuigan et al. (2006). An impedimetric sensor for monitoring the growth of staphylococcus epidermidis RP62A. In A.P. Dhawan & M. Akay (Eds.), *28th Annual International Conference of the IEEE-EMBS*, IEEE, New York (p. 535).

Owens, J.D. (1985). Formulation of culture media for conductimetric assays: Theoretical considerations. *Journal of general microbiology, 131,* 3055-3076.

Poulsen, L.V. (1999). Microbial biofilm in food processing. *LWT-Food Science and Technology, 32,* 321-326.

Stern, O. (1924). The theory of the electrolytic double layer. *Zeit Electrochemistry, 30,* 508-516.

Stewart, G.N. (1898). The changes produced by the growth of bacteria in the molecular concentration and electrical conductivity of culture media. *Journal of Experimental Medicine, 4,* 235-243.

Trotman, R.E. (1978). Mechanizing microbiology: The advantages and future prospects. In A.N. Sharpe & D.S. Clarke (Eds.), *Mechanizing microbiology* (pp. 3-9). SpringField, UK.

Ur, M., & Brown, F.J. (1975). Impedance monitoring of bacterial activity. *Journal of Medical Microbiology, 8,* 19-27.

Yang, L., Li, Y., Griffis, C.L., & Johnson, M.G. (2004). Interdigitated microelectrode (IME) impedance sensor for the detection of viable Salmonella typhimurium. *Biosensors and Bioelectronics, 19,* 1139-1147.

Zoltowski, P. (1998). On the electrical capacitance of interfaces exhibiting constant phase element behaviour. *Journal of Electroanalytical Chemistry, 443*, 149-154.

KEY TERMS

Biofilm: Microbial biofilms are defined as a matrix-enclosed bacterial population adhered to a surface or each other. An organic matrix protects these sessile organisms. Biofilms are characterized in part by the production of a network of highly hydrated extracellular polysaccharides (EPS).

Charge Transfer Resistance (R_{CT}): The DC current that manages to cross the electrode-electrolyte interface experiences a charge transfer resistance. The expression for R_{CT} is derived from the Butler-Volmer equation where:

$$R_{CT} = \frac{RT}{nF}\frac{1}{i_0}$$

R is the gas constant, T is temperature (K), F is the Faraday constant, n is the number of electrons involved in the electrode reactions, and i_0 is the exchange current density.

Constant Phase Element: In order to facilitate equivalent circuit modelling, the ideal capacitance is often substituted by an empirical constant phase element (CPE). The CPE is not physically realisable with ordinary lumped electric components but it is usually described as a capacitance that is frequency dependent. The use of a CPE reflects the nonhomogeneity of the system under investigation.

Hospital-Acquired Infections (HAI): Infections that are neither present nor incubating when a patient enters the hospital. Effects vary from discomfort to prolonged or permanent disability where HAI's may contribute directly or substantially to a patient's death.

Impedance: The opposition to flow of an alternating current. Impedance is a complex entity that describes all resistance to current flow. Impedance is composed of resistance (the real component) and capacitive and or inductive reactance (the imaginary components).

Solution Resistance (R_s): The opposition to current flow through the solution/electrolyte by the migration of ions. In general, the solution resistance is considered to decrease with increase in microbial growth.

Electrocutaneous Stimulation of Skin Mechanoreceptors for Tactile Studies with Functional Magnetic Resonance Imaging

Valentina Hartwig
University of Pisa, Italy
C.N.R. Institute of Clinical Physiology, Italy

Claudia Cappelli
University of Pisa, Italy

Nicola Vanello
University of Pisa, Italy
C.N.R. Institute of Clinical Physiology, Italy

Emiliano Ricciardi
University of Pisa, Italy

Enzo Pasquale Scilingo
University of Pisa, Italy

Giulio Giovannetti
C.N.R. Institute of Clinical Physiology, Italy

Maria Filomena Santarelli
C.N.R. Institute of Clinical Physiology, Italy

Vincenzo Positano
C.N.R. Institute of Clinical Physiology, Italy

Pietro Pietrini
University of Pisa, Italy

Luigi Landini
University of Pisa, Italy

Antonio Bicchi
University of Pisa, Italy

INTRODUCTION

In many simulation and exploration tasks such as exploring the real and virtual environment, tactile information is necessary to get surface information of objects. Moreover, in rehabilitation and sensory prostheses training, this kind of sensory information is indispensable (Yarimaga, Lee, Lee, & Ryu, 2005).

Since the 1980s, electrotactile stimulation has been considered a possible way to reproduce the glabours skin tactile sensations, first by using intracutaneous electrodes inserted into the nerves and then by using transcutaneous electrodes placed in contact with the skin.

Electrotactile stimulation is a potentially useful method for sensory augmentation or substitution and permits accurately controlling the perceived stimulus intensity that is fundamental in any tactile communication system. In several virtual reality applications, the electrostimulation could be a viable solution in substituting for tactile displays, increasing the augmented reality performance.

If it could be possible to elicit selectively each kind of skin afferent fibers using analogously to mechanical stimulus, any sensation could be evocated by combining specific inputs.

To evoke specific sensations related to specific types of *mechanoreceptors,* we consider here the possibility of using electrotactile stimulation, varying the amplitude and the frequency of the stimulating electric signal, but maintaining the same waveform. According to a different specificity of the mechanoreceptors at different mechanical stimulus (Kaczmarek, Webster, Bach-y-Rita, & Tompkins, 1991), there might exist receptor specificity for different amplitude and frequency of electrical stimulus. In this case it could be possible to stimulate specifically several types of mechanoreceptors maintaining the same waveform and varying only its amplitude or frequency, with the possibility of creating a different sensation for tactile studies and augmented reality applications.

BACKGROUND

The *microstimulation* technique of tactile receptors by using electric current has been largely used to study the specific role of several types of mechanoreceptors and characterize their functional properties (Vallbo, Olsson, Westberg, & Clark, 1984). Passing small electric currents through microneurography electrodes placed directly on the nerve ending of the receptors, it is possible to evoke several localized sensations such as flutter or pressure.

Many authors have proposed several techniques to record this sensation in awake human: Trulsson Francis, Kelly, Westling, Bowtell, and McGlone (2001) demonstrated that intraneuronal microstimulation of single afferents produces robust hemodynamic responses in somatosensory cortex that can be measured using the *functional Magnetic Resonance Imaging* (fMRI) technique.

Another technique of electrical stimulation of skin receptors is known as "electrocutaneous stimulation": with the term electrocutaneous (or electrotactile) is intended the evocation of a tactile sensation using an electric current flowing through the skin, via electrodes placed on the skin surface. A device that stimulates nerve afferents within the skin by electric current is known as "electrocutaneous display" and can be constituted by several surface electrodes.

Because their small size, longer durability, energy efficient, and ease of use, electrocutaneous displays are superior to mechanical tactile displays.

Several papers report the use of this technique in sensory substitution system for blind or deaf persons (Bak, Girvin, Hambrecht, Kufta, Loeb, & Schmidt, 1991; Eisenberg, Maltan, Portillo, Mobley, & House, 1987) and in training for prostheses use. Electrocutaneous stimulation has been also used in augmented reality and telepresence in order to provide the user with tactile information (Nojima, Sekiguchi, Inami, & Tachin, 2002).

The mechanism of tactile stimulation was first described in literature approximately in 1960 and the first tactile display was proposed in 1970 (Strong & Troxel, 1970). It was based on the principle that an electric current pulse from surface electrodes generates an electric field inside the skin, which induces nerve activity.

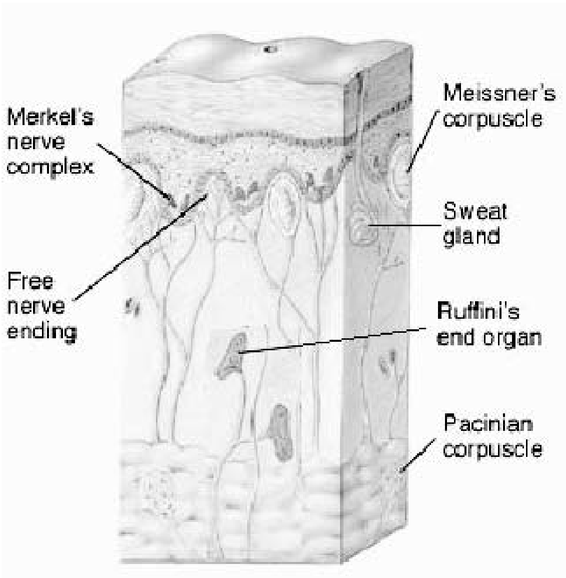
Approximately 44% of mechanoreceptors lying into the human hand are found to be slowly adapting (SA) (i.e., they also respond with a sustained discharge to static tissue deformation), while the remaining are fast adapting (FA), only responding to the rate of skin indentation and its higher derivatives. Depending on the extension of their receptive fields, SA and FA tactile units can be subdivided into two categories: type I have restricted and sharply defined receptive fields and type II have larger fields and less precise contours. The correspondence between SAI and *Merkel's complexes,* FAI and *Meissner's corpuscles,* SAII and *Ruffini's endings*, and FAII and *Pacinian corpuscles* is widely accepted (Johnson, 2001).

Each class of mechanoreceptors responds to skin deformation and motion in a different way (Kaczmarek et al., 1991). If it could be possible to find the electric current which is able to elicit selectively each kind of afferent fibers analogously to mechanical stimulus, any sensation could be evocated by combining specific inputs.

Kajimoto, Hawakami, Maeda, and Tachi (2002) have already shown that electrical selective stimulation is possible using anodic or cathodic current: they called the specific stimuli "tactile primary colours" in analogy to the visual system and its primary colours (red, green, and blue). In their paper the authors show that Meissner corpuscles can be selectively stimulated in

E

Figure 1. Skin mechanoreceptors



a specific setup of electrical stimulation but Pacinian corpuscles can not be selectively stimulated because the electrical signal inevitably co-activated the shallow parts also.

To evoke specific sensations related to specific types of mechanoreceptors, we considered the possibility of using electrotactile stimulation varying the amplitude and the frequency of the stimulating electric signal, but maintaining the same waveform. We only focus on Merkel and Meissner corpuscles which are at less depth in the skin respect to Pacinian and Ruffini afferents.

In order to verify the evocation of a tactile sensation by means of the electrical signal, it is possible to use the fMRI technique. This technique allows localization of blood flow variations and oxygen consumption due to neuronal firing, then it is possible to visualize the active brain regions during perception and processing of a sensory impulse from a given body district (Hammeke & Yetkin, 1994).

DESIGN OF AN ELECTROCUTANEOUS DISPLAY

Safety Criteria

As the microstimulation is performed by an electric device, *safety criteria* for electric safety, in terms of high voltage and pulsed current, were considered and studied.

Predicting the current flow in the skin is very difficult because of the many factors which influence it, for example, whether the body skin is moist or dry, the contact area, and pressure. A good review of physical basis for the design of efficacious and safe protocols for electrical stimulation is in Merrill, Bikson, and Jefferys (2005).

Moreover, electrochemical reaction could be occurring at the electrode interface and then they have to be considered for the choose of electrode material (Brummer & Turner, 1977).

The design of the electric microstimulator was developed in accordance to results from *MR safety and compatibility* studies. The observance of these criteria is a crucial aspect in view of performing specific experiments inside the MR scanner during a fMRI acquisition of brain functional activation.

Before introducing any device into the MR suite, MR safety and compatibility of the device and its operation in the magnetic field should be carefully verified according to systematic rules and procedures (Shellock, 1998). Hence, in order to design a new device which allows tactile stimuli inside the MRI scanner environment during fMRI studies, definitions and classifications regarding MR safety and compatibility and the interaction between MRI primary components and mechatronic devices have been considered. The Food and Drug Administration (FDA) document (1997) on medical device interactions with Magnetic Resonance Imaging systems reports the most descriptive definitions and classifications regarding MR safety and compatibility that should be used for experimental protocols of compatibility evaluation.

Methods

The stimulator device is supplied by low power batteries and is able to generate high voltage pulses; it is equipped with two independent channels for stimulation of both hands. By means of a couple of potentiometers for each channel, it is possible to set the output voltage amplitude and frequency values, in order to vary the kind of the stimulus and obtain different sensations due to specific classes of stimulated mechanoreceptors.

The output voltage ranges from 0 to 450 V, while frequency varies between 90 and 110 Hz; however because of the very high skin impedance, the electric

Figure 2. Photo of designed electrical stimulator internal

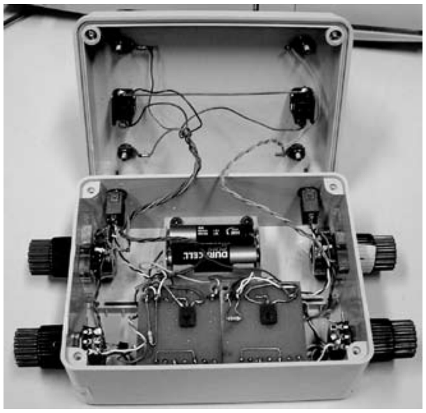
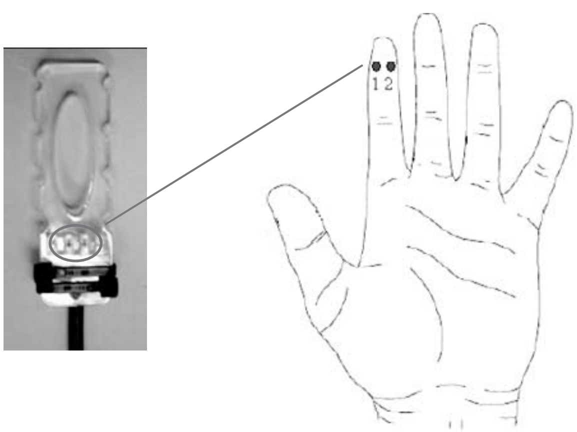


Figure 3. Stimulator pad with two aluminum electrodes and their position on the index fingertip



current flowing is 1-2 mA. The output waveform is a pulsed decreasing exponential wave. This permits to reduce the Joule effect produced by the electric current flow through the skin.

According to the electric safety criteria the stimulator device is floating and supplied by batteries.

According to the MR safety and compatibility criteria, the part which should be placed within the scanner room does not present metallic materials. For this reason, the stimulator pad is created choosing MR-compatible and nontoxic materials, Plexiglass and Aluminum, because it has to be placed on the subject hand inside the scanner room. Two 1mm x 1mm Aluminum square electrodes, at distance of 5 mm, are fixed on the pad.

Coaxial cables are used to connect the two stimulator electrodes inside the MR scanner room with the device inside the console room in order to minimize RF interferences.

Device Set Up

Before testing the device during a fMRI acquisition, we performed several psychophysical experiments in laboratory. A group of subjects, stimulated by an electric current which was varied in amplitude and frequency, was asked to report the perceived tactile sensation. Specifically, starting from the lowest amplitude and highest frequency, the electric current was firstly increased in intensity up to subjects felt some sensation. In 95% of the cases, the sensation was defined like a slipping, vibration sensation.

Afterwards, we decreased the frequency of the electric current down to subjects were able to feel a different sensation. In 90% of the cases, this second sensation was defined like a pin-prick sensation.

According to Johnson (2001), *Meissner* corpuscles (RA1 afferents) are very sensitive to dynamic skin deformation and they transmit a robust neural image of skin motion. For this reason, we can hypothesize that the specific value of amplitude and frequency used in the first case, in which subjects perceived slipping sensation, could elicit RA1 receptors.

Instead, *Merkel* corpuscles (SA1 afferents) are more sensitive to points, edges, and curvature, which is a consequence of their selective sensitivity to strain energy density. They are responsible for shape, orientation, and texture perception. Therefore, we can conjecture that the specific value of amplitude and frequency of the electric current achieved in the case where subjects felt pin-prick sensation could be responsible for eliciting SA1 receptors.

MR Safety and Compatibility Test

The designed device was examined in a MR environment (Signa Horizon 1.5T, GE Medical Systems) during image acquisition in order to test its MR safety and compatibility.

We placed the device inside the console room and the stimulator electrode inside the scanner room on the patient bed in the right position for the stimulation task on a patient (the cable that connects the stimulator electrode to the device passed through the wire guides of the MR shield).

During the test we scanned a spherical phantom of $CuSO_4$ solution, using a GE-EPI (gradient echo, eco planar imaging) with the following parameters: TE/TR 40/3000 msec, bandwidth 62.5 kHz, FOV 24 cm, resolution 64x64 pixels, Flip angle 90°, Slice thickness 5 mm, number of slices 25, 25 volumes acquired. This is a common use sequence for fMRI studies. After the images acquisition the signal to noise ratio (SNR) was calculated (Sijbers, Den Dekker, Van Audekerke, Verhoye, & Van Dyck, 1998). As a second parameter we estimated the standard deviation of the image intensity time course (SD) in each image sequence.

We looked for differences between image sets acquired with no device (reference images) and the image sets acquired with the electrode placed on the patient bed in several conditions: no electric current and electric current with different amplitude and frequency for different sensation stimulations. In order to evaluate the device compatibility with the MR environment, it is not sufficient to compare the two sets of images because several artifacts are not visible, so we used statistical t and z tests to detect parameter differences (Gassert et al., 2006). The analysis of these images and the application of the statistical tests to the images confirms the compatibility in MR-environment and the absence of unexpected and dangerous effects.

fMRI Acquisition on a Subject

After the evaluation of MR safety and compatibility of the device, we performed a pilot fMRI study on one subject.

The subject was a female 27 years old: she had been lying down on the sliding bed and the operator placed the birdcage head coil around her head. The stimulator electrode was placed on the subject's left index finger and fixed in order to avoid relative movements.

First of all we acquired the anatomic image of the subject brain and subsequently we performed a functional study using a classical fMRI sequence (GE-EPI: TE/TR 40/3000 msec, bandwidth 62.5 kHz, FOV 24 cm, resolution 64x64 pixels, Flip angle 90°, Slice thickness 5 mm, number of slices 25, 25 volumes acquired).

In order to explore brain activity during electric stimulation by using fMRI, we used a block design approach with a stimulation of 10 seconds, an interstimulus interval of 20 seconds, an initial and a final baseline condition of 25 seconds each (total number of stimulation interval=8). The choice of this experiment block designed was guided by safety criteria (for example the Joule effect presence) and analysis criteria (the stimulus intervals have to be short to avoid saturation effects on the receptors).

The fMRI acquisition involved the following runs:

- ***Run 1*****:** stimulation with specific value of amplitude and frequency of electrical signal in order to obtain a relative slipping sensation between the electrode and the subject finger.
- ***Run 2*****:** subject finger stimulation moving, by an operator in the scanner room, a piece of velvet in order to obtain a functional activation of a real tactile sensation.
- ***Run 3*****:** stimulation with specific value of amplitude and frequency of electrical signal in order to obtain a pin prick sensation on the subject finger.

Before the acquisition, a calibration of the device was made asking to the subject what kind of sensation she perceived.

Experimental Results

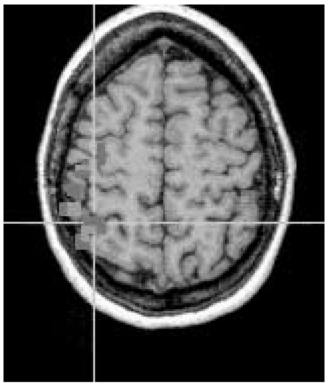
For every one of the three stimulation runs performed, a robust hemodynamic response was observed. Figure 4 shows the results for run 1, which are typical of all data acquired.

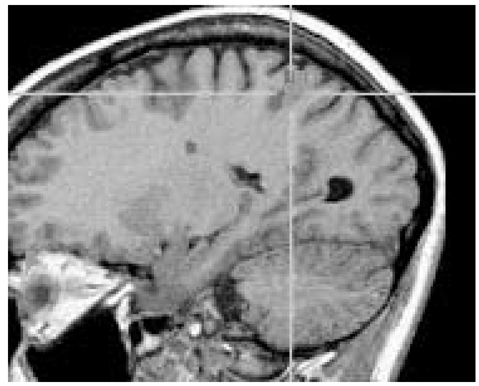
In correspondence of the primary sensorimotor cortex (SI) area (Hammeke & Yetkin, 1994), the statistical analysis, performed using AFNI (Cox, 1996), shows a neuronal activation; this confirms the hypothesis that the electric current caused a tactile sensation.

The graph in Figure 5 shows the correlation between the stimulation block and the intensity level for an image pixel that corresponds to the hemodynamic response: it is possible to note the good correlation between the stimulus intervals (red line, high intervals) and the activation.

Low frequency fluctuations are usually observed in fMRI time series. These may be related to long range physiological changes or slow movements of the subjects' head. For acquisition lengths of some minutes, these changes can be modelled with a linear ramp. The raw data in Figure 5 show an offset linear with time. Neural activity related to signal changes are small with respect to the baseline level and may be hidden by

Figure 4. Cortical activation during run 1 (slipping sensation). (a) axial view, (b) sagittal view


a)


b)

noise. In this task, even electrode displacement due to subject finger movement may occur. For this reasons, it is difficult to judge signal behaviour related to each single event, and the correlation between the entire time series and the stimulus paradigm must be evaluated.

These results demonstrate unequivocally that the electric stimulation really produces a robust hemodynamic response in the SI cortex, where touch resides, and it can be measured using fMRI.

FUTURE TRENDS

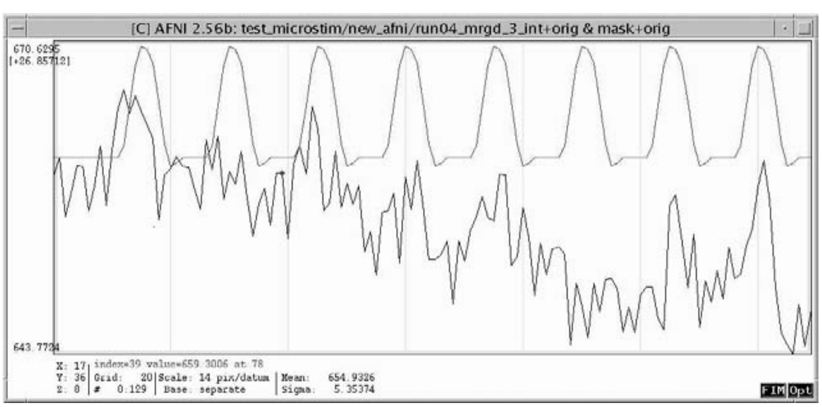
Future developments will concern a more accurate discrimination of the specific sensations reproduced using the electric current and major efforts will be focused in finding a reliable relationship between mechanical and electric stimuli eliciting mechanoreceptors. It would be interesting to stimulate different locations in order to study how different the elicited patterns are.

CONCLUSION

The goal of our work was to investigate the possibility of stimulating selectively different mechanoreceptors to evoke specific tactile sensations by means of injection of electric current through the finger skin. Possible fields of application are fMRI tactile studies or augmented reality technology. In this preliminary work, we described the first prototype of an electric stimulator, verifying the MR safety and compatibility, and performed a preliminary experiment with a subject under fMRI exam.

In the device design the electric and magnetic safety criteria was followed in order to avoid every

Figure 5. Activated area pixel intensity (bottom black line) and electrical stimulus (up red line) time evolution



dangerous effect due to the use of electric current in the human body.

The MR safety and compatibility of the prototype has been demonstrated using a statistical test to the localization of eventual artefacts in the MR images. Finally, the prototype was tested on a subject during a fMRI examination in order to analyze the functional brain activation due to the tactile sensation caused by the electric current through the electrode. The AFNI analysis showed the effective activation of the primary and secondary sensorimotor areas in which the touch resides, and a good correlation between the stimulation block paradigm and the brain specific areas activity. These preliminary results demonstrate that electrocutaneous stimulation performed by means of our device produces robust and focal hemodynamic responses in somatosensory cortex that can be measured by fMRI technique.

Future developments will concern a more accurate discrimination of the specific sensations reproduced using the electric current and major efforts will be focused in finding a reliable relationship between mechanical and electric stimuli eliciting mechanoreceptors. It would be interesting to stimulate different locations in order to study how different the elicited patterns are.

REFERENCES

Bak, M., Girvin, J.P., Hambrecht, F.T., Kufta, C.V., Loeb, G.E., & Schmidt, E.M. (1991). Visual sensations produced by intracortical microstimulation of the human occipital cortex. *Med. Biol. Eng. Comput., 28*, 257-259.

Brummer, S., & Turner, M. Electrochemical considerations for safe electrical stimulation of nervous system with platinum electrodes. *IEEE Ttransactions on Biomedical Engineering, 10,* 60-62.

Cox, RW. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res., 29*, 162-173.

Eisenberg, L.S., Maltan, A.A., Portillo, F., Mobley, J.P., & House, W.F. (1987). Electrical stimulation of the auditory brainstem in deafened adults. *J. Rehab. Res. Develop., 24*, 9-22.

Gassert, R., Vanello, N., Chapuis, D., Hartwig, V., Scilingo, E.P., Bicchi, A., et al. (2006). Active mechatronic interface for haptic perception studies with functional magnetic resonance imaging: compatibility and design criteria. In *Proceedings of IEEE ICRA 2006* (pp. 3832-3837).

Hammeke, T.A., & Yetkin, FZ. (1994). Functional magnetic resonance imaging of somatosensory stimulation. *Neurosurgery, 35,* 677-681.

Johnson, KO. (2001). The roles and functions of cutaneous mechanoreceptors. *Current Opinion in Neurobiology, 11*, 455-461.

Kaczmarek, K.A., Webster, J.G., Bach-y-Rita, P., & Tompkins, W.J. (1991). Electrotactile and vibrotactile displays for substitution systems. *IEEE Ttrans. on Biomed. Eng., 38*, 1-16.

Kajimoto, H., Kawakami, N., Maeda, T., & Tachi, S. (2202). Electrocutaneous display with receptor selective stimulations. *Electronics and Communications in Japan, Part 2, 85*, 120-128.

Merrill, D.R., Bikson, M., & Jefferys, J.G.R. (2005). Electrical stimulation of excitable tissue: Design of efficacious and safe protocols. *Journal of Neuroscience Methods, 141*, 171-198.

Nojima, T., Sekiguchi, D., Inami, M., & Tachin, S. (2002). The smarttool: A system for augmented reality of haptics. In *Proc. IEEE VR'2002* (pp. 67-72).

Shellock, F.G. (1998). *Pocket guide to MR procedures and metallic objects: Update 1998.* Philadelphia, PA: Lippincott-Raven Pub..

Sijbers, J., Den Dekker, A.J., Van Audekerke, J., Verhoye, M., & Van Dyck, D. (1998). Estimation of the noise in magnitude MR images. *Magnetic Resonance Imaging, 16*(1), 87-90.

Strong, R.M., & Troxel, D.E. (1970). An electrotactile display. *IEEE Trans. Man-Mach. Sys., MMS-11*, 72-79.

Trulsson, M., Francis, S.T., Kelly, E.F., Westling, G., Bowtell, R., & McGlone, F. (2001). Cortical responses to single mechanoreceptive afferent microstimulation revealed with fMRI. *NeuroImage, 13*(4), 613-622.

U.S. Department of Health and Human Services, Food and Drug Administration, Center for Devices and Radiologicial Health (Eds.). Magnetic resonanceworking group draft. A primer on medical device interactions with magnetic resonance imaging systems. Retrieved February 12, 2008, from http://www. fda. gov/cdrh/ode/primerf6.html

Vallbo, B., Olsson, K.A., Westberg, K.G., & Clark, F.J. (1984). Microstimulation of single tactile afferents from the human hand. *Brain, 107*, 727-749.

Yarimaga, O., Lee, J., Lee, B.C., & Ryu, J. (2005, June 2-5). Tactile sensation display with electrotactile interface. In *Proocedings of ICCAS2005, KINTEX*, Gyeonggi-Do, Korea.

Key Terms

Electrocutaneous (or electrotactile) Stimulation: The evocation of a tactile sensation using an electric current flowing through the skin, via electrodes placed on the skin surface.

Electrocutaneous Display: A device that stimulates nerve afferents within the skin by electric current and can be constituted by several surface electrodes.

Functional Magnetic Resonance Imaging (fMRI): Imaging technique based on blood oxygenation level dependent (BOLD) effect that allows localization of blood flow variations and oxygen consumption due to neuronal firing.

Intracutaneous Electrodes: Electrodes directly inserted into the nerves.

Mechanoreceptors: Cutaneous receptors that are sensitive to mechanical stimuli that are skin motion, skin stretch, pressure, vibration, points, edges, curvature.

Meissner Corpuscles: Rapidly adapting (RA) afferents that innervate the skin densely and are sensitive to dynamic skin deformation. They transmit a robust neural image of skin motion.

Merkel Cells: Slowly adapting (SA) afferents that are sensitive to points, edges, and curvature. They are responsible for form and texture perception.

Transcutaneous Electrodes: Electrodes placed in contact with the skin.

Electronic Test Management Systems and Hospital Pathology Laboratory Services

Andrew Georgiou
The University of Sydney, Australia

Johanna Westbrook
The University of Sydney, Australia

Joanne Callen
The University of Sydney, Australia

Jeffrey Braithwaite
University of New South Wales, Australia

INTRODUCTION

Pathology can be described as the branch of medicine that deals with the nature, causes, and process of disease (McGrath, 2003). Pathology laboratories consider clinical and pathologic data and integrate them within an ever-changing context and then transmit a meaningful answer back to doctors and patients. In doing so, pathology laboratories play a key role in translating data into meaningful information (Hardwick, 1998). Pathology services are information intense organisational bodies that rely heavily on the proficient administration of information for patient care purposes (Travers, 1997). It is estimated that 70% of all important decisions affecting a patient's life involve a laboratory or pathology test, and pathology data represent an average of 70% of documents residing in electronic repositories (Becich, 2000).

Yet, pathology services are still widely seen as a backroom function with many people unaware of their importance. Pathology has been dubbed the "hidden science that saves lives" by the Royal College of Pathologists in England (The Royal College of Pathologists, 2000). Pathology departments are facing challenges from new information and communication technology (ICT) advances and the advent of managed care approaches to health care planning and delivery. The Review of NHS Pathology Services in England in 2006 emphasised the key role of pathology services in patient pathways that begins with the choice of the most suitable test or investigation, and proceeds to the interpretation and supply of clinical advice across clinical specialties (Review of NHS Pathology Services in England, 2006). ICT developments are behind many of the moves aimed at extending the role of pathology services beyond the basic request and reporting cycle (Friedman, 1996).

BACKGROUND

In the last 10 years, there has been much emphasis on the potential for computerised provider order entry (CPOE) systems to improve the provision and quality of health care (Doolan & Bates, 2002; Sittig & Stead, 1994). CPOE systems provide clinicians with the ability to place orders directly into computers linked to databases containing specific clinical information and decision-support software (Birkmeyer, Lee, Bates, & Birkmeyer, 2002). Many health care systems internationally are involved in the implementation of CPOE systems (Humber, 2004; NSW Government Action Plan for Health, 2002; The Leapfrog Group for Patient Safety, 2003). These systems are cornerstones for the establishment of electronic medical records (Hwang, Park, & Bakken, 2002).

Even though there has been substantial support for the implementation of CPOE systems (The Leapfrog Group for Patient Safety, 2003) along with a growing evidence base of their impact on the delivery of health care (Birkmeyer et al., 2002; Doolan & Bates, 2002) and its efficiency (Mekhjian et al., 2002), uptake has been neither rapid nor even (Ash, Gorman, Seshaddri, & Hersh, 2004). Some of the initial enthusiasm for

CPOE systems has been tempered by high profile cases of physician resistance (Berger, 2004), and implementation difficulties (Dykstra, 2002) along with concern about the huge investment and costs involved (Ash & Bates, 2005). Moreover, evidence about the unintended consequences of CPOE systems (Ash, Berg, & Coiera, 2004; Campbell, Sittig, Ash, Guappone, & Dykstra, 2006) and their potential to facilitate new types of errors (Koppel et al., 2005) have led to a renewed focus on the importance of evaluation (Ammenwerth & de Keizer, 2005; Friedman & Wyatt, 1997; Gell, 2001) as a means to improve their design and implementation.

So far the attention of the research and evaluation literature has tended to focus on high profile issues like medication errors, with less attention to areas like pathology laboratories and medical imaging, which together make up a major proportion of hospital orders (Abelson, Connelly, Klee, Maag, & Smith, 2001; Georgiou, Williamson, Westbrook, & Ray, 2007). CPOE is by definition a system-wide phenomenon with implications for the way the whole hospital and related entities work and function. These issues and challenges cannot be addressed by silo-based approaches where departments are considered independently of each other (Georgiou & Westbrook, 2006; Stablein et al., 2003). Pathology services are themselves made up of a number of organisational subparts each with their own ways of operating and functioning (Davidson & Chismar, 1999b), that will be affected by (and in turn affect) CPOE implementation (Wears & Berg, 2005). In the following sections, we draw on existing research evidence and literature reviews (Georgiou & Westbrook, 2006; Georgiou et al., 2007) alongside our own research experience to formalise an evaluation framework that can be used to assess the impact of CPOE on pathology services.

EVALUATING THE IMPACT OF CPOE ON PATHOLOGY PROCESSES

A systematic review by Georgiou et al. (2007) conceptualised three stages in the pathology test management process beginning with: (a) the decision of the doctor or responsible clinician (doctor or other delegated health professional) to order a pathology test; followed by (b) the processing of the test order in the pathology laboratory and ending with (c) a result that is communicated to the clinician and health care team responsible for the care of the patient, which will then be used as part of the clinical decision-making process (Georgiou et al., 2007). Each of these stages involves a dimension of time (Howanitz & Howanitz, 2001) which can be measured by turnaround time (TAT) indicators involving a number of measures including: (1) Laboratory TAT - the time taken for the test order to be processed in the laboratory before a result is issued, and (2) Total TAT—the total time it takes for an order to be placed, processed and a result issued (Georgiou et al., 2007).

Test Order Stage

Each of the stages in the pathology test order process can be assessed with a range of indicators that have been used to monitor the impact of CPOE systems on pathology services and patient care (Georgiou et al., 2007). The ability of CPOE systems to provide decision support will most likely have an effect on the first stage of the pathology test order process involving the clinician's decision about which test to order. Some researchers have paid particular attention to the ability of decision support systems to affect clinical compliance with practice guidelines (Overhage, Tierney, Zhou, & McDonald, 1997; Solomon et al., 1999). Decision support may also affect the appropriateness and volume of tests ordered. This is particularly the case for "redundant" tests, that is, tests that are repeated within an inappropriate time frame and provide no additional information (Bates et al., 1998; van Walraven & Naylor, 1998). The volume of tests can in turn be measured in different ways, for example, the number of tests per day (Hwang et al., 2002), or for a specified period, or per patient/admission (Tierney, Miller, & McDonald, 1990; Westbrook, Georgiou, Dimos, & Germanos, 2006). The volume of tests is likely to have a significant effect on test costs which can also be measured in various ways such as: total laboratory costs (Nightingale, Peters, Mutimer, & Neuberger, 1994) or per admission (Tierney et al., 1990). Some research has concentrated on the effect that CPOE systems have on work practices of clinicians and pathology services staff. One of the most important issues in this area involves quantifying the time spent ordering tests and its impact on other tasks (Shu et al., 2001). Another key concern in the area of work practices is ensuring that the new technology does not foster practices which affect the quality and safety of the ordering process (Koppel et al., 2005).

Test Processing Stage

Few studies have looked at the impact of CPOE on the pathology test processing stage (Georgiou et al., 2007). In part, this may be because existing laboratory quality control processes are used to ensure the accuracy and reproducibility of results (Tetrault, 2001), and CPOE is not expected to greatly impact this area. The greatest number of errors that occur during this stage are a result of incorrect transcriptions and specimen errors (Bonini, Plebani, Ceriotti, & Rubboli, 2002; Plebani & Carraro, 1997). CPOE is expected to affect this area because it provides a template for accurate and legible ordering by clinicians and eliminates the need for laboratories to record orders (Georgiou et al., 2007). One method that has been used to monitor the impact of CPOE systems in this area is to record the level of telephone activity between wards and the laboratory as a means of ascertaining whether computerised test management systems reduce the need to chase up missing or unclear documentation details (Ostbye, Moen, Erikssen, & Hurlen, 1997).

Test Result Application

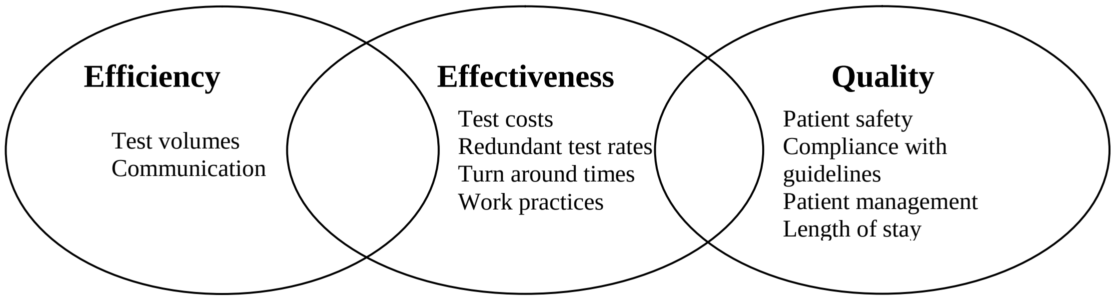
The final stage of the process involves the application of test results as part of delivery of patient care. This is the stage where the test process impacts directly on the outcomes and quality of patient care (Nevalainen et al., 2000). Evaluation of the role of CPOE in this stage of the test management process can include comparison of the time it takes to reach a diagnosis (Smith & McNeely, 1999) or to initiate treatment (Kuperman et al., 1999) when using a CPOE system. It is also possible to monitor the length of stay of patients (Neilson et al., 2004) and aspects of patient safety such as adverse events and mortality (Kuperman et al., 1999). CPOE systems have been shown to be successful at improving the speed with which test results are made available to clinicians. However, little research attention has been given to how individual clinicians are able to manage high volume and rapid transfer of clinical information (Kilpatrick & Holding, 2001; Murff et al., 2003; Poon et al., 2004).

FUTURE TRENDS: EVALUATION FRAMEWORK

The various measures of CPOE performance and impact provide a framework (see Figure 1) with which to assess (a) efficiency (value and efficacy of services in terms of cost, time and standards of practice) (Potter, 2000; Scriven, 1991); (b) effectiveness (the best possible outcome) (Potter, 2000) or success of the intervention (Scriven, 1991); and (c) quality (ensuring that the right thing is performed well (Brook & Kosecoff, 1988; Donabedian, 1988) and that the system meets identified needs and other relevant standards (Davidson, 2005)). But as the interconnected components of Figure 1 illustrate, the domains of efficiency, effectiveness and quality are not mutually exclusive and involve measures that clearly straddle domains (e.g., length of stay). One of the limitations of existing literature in this field is that evaluations are often based on indicators measured in isolation and disconnected from each other. It is important to maintain a holistic overview of indicator measurements, understanding that the net effect of any particular information system will consist of a balance of positives and negatives, and possibly successes and failures (Pawson, 2004).

The framework outlined in Figure 1 concentrates on quantitative measures of evaluation. In recent years there has been a growing number of researchers (Ammenwerth et al., 2004; Ash, Sittig, Seshadri, Dykstra, Carpenter, & Starvi, 2004; Greatbatch, Murphy, & Dingwall, 2001; Kaplan, 2001; Stoop & Berg, 2003) who have employed qualitative research methods to evaluate health information systems, including CPOE systems in pathology (Callen, Westbrook, & Braithwaite, 2006; Davidson & Chismar, 1999a; Georgiou, Westbrook, Braithwaite, & Iedema, 2005; Georgiou, Westbrook, Braithwaite et al., 2007). These approaches have contributed to a better understanding of the context (e.g., effect on the hospital and clinical environment) (Callen, Braithwaite, & Westbrook, 2006) and processes of ICT implementation including their impact on communication channels between departments and disciplines (Aydin, 1989; Dykstra, 2002). Together with quantitative research, they can add a valuable multimethod dimension to evaluation studies (Georgiou, Westbrook, Braithwaite, Iedema, Dimos, & Germanos, 2005; Westbrook, Braithwaite, Iedema, & Coiera, 2004).

Figure 1. Framework for assessing the impact of CPOE on pathology services



CONCLUSION

The implementation of CPOE systems is increasing internationally and there is a strong evidence base about their ability to contribute to the efficiency and effectiveness of health care delivery. However, this evidence has tended to be concentrated in a small number of United States' hospitals (Chaudhry et al., 2006; Shekelle, Morton, & Keeler, 2006). This means that the generalisability of this evidence is limited because these hospitals often have home grown systems that are not commercially available and which evolved over many years, mostly in academic teaching hospitals (Classen, Avery, & Bates, 2007). This has underscored the drive to extend and systematise the evaluation of CPOE systems across a range of diverse settings (Classen et al., 2007) including pathology services. Evaluation frameworks built upon existing evidence and utilising the experience of skilled practitioners and researchers can help to orient the evaluation process and provide comparative and generalisable data with which to optimise the effect of CPOE systems on patient care delivery.

REFERENCES

Abelson, D., Connelly, D., Klee, G., Maag, T., & Smith, J. (2001). *Technology assessment: Computerized physician order entry: Institute for clinical systems improvement.* Retrieved March 25, 2008, from www.icsi.org

Ammenwerth, E., Brender, J., Nykanen, P., Prokosch, H. U., Rigby, M., Talmon, J., et al. (2004). Visions and strategies to improve evaluation of health information systems. Reflections and lessons based on the HIS-EVAL workshop in Innsbruck. International *Journal of Medical Informatics, 73*(6), 479-491.

Ammenwerth, E., & de Keizer, N. (2005). An inventory of evaluation studies of information technology in health care trends in evaluation research 1982-2002. *Methods of Information in Medicine, 44*(1), 44-56.

Ash, J. S., & Bates, D. W. (2005). Factors and forces affecting EHR system adoption: Report of a 2004 ACMI discussion. *Journal of the American Medical Informatics Association, 12*(1), 8-12.

Ash, J. S., Berg, M., & Coiera, E. (2004). Some unintended consequences of information technology in health care: The nature of patient care information system-related errors. *Journal of the American Medical Informatics Association, 11*(2), 104-112.

Ash, J. S., Gorman, P. N., Seshaddri, V., & Hersh, W. R. (2004). Computerized physician order entry in U.S. hospitals: Results of a 2002 survey. *Journal of the American Medical Informatics Association, 11*(2), 95-99.

Ash, J. S., Sittig, D. F., Seshadri, V., Dykstra, R. H., Carpenter, J. D., & Starvi, P. Z. (2004). Adding insight: A qualitative cross-site study of physician order entry. *Medinfo., 11*(Pt 2), 1013-1017.

Aydin, C. E. (1989). Occupational adaptation to computerized medical information systems. *Journal of Health and Social Behavior, 30*(2), 163-179.

Bates, D. W., Boyle, D. L., Rittenberg, E., Kuperman, G. J., Ma'Luf, N., Menkin, V., et al. (1998). What proportion of common diagnostic tests appear redundant? *American Journal of Medicine, 104*(4), 361-368.

Becich, M. J. (2000). Information management: moving from test results to clinical information. *Clinical Leadership & Management Review, 14*(6), 296-300.

Berger, R., Kichak BA. (2004). Computerized physician order entry: Helpful or harmful? *Journal of the American Medical Informatics Association, 11,* 100-103.

Birkmeyer, C. M., Lee, J., Bates, D. W., & Birkmeyer, J. D. (2002). Will electronic order entry reduce health care costs? *Effective Clinical Practice, 5*(2), 67-74.

Bonini, P., Plebani, M., Ceriotti, F., & Rubboli, F. (2002). Errors in laboratory medicine. *Clinical Chemistry, 48*(5), 691-698.

Brook, R., & Kosecoff, J. (1988). Commentary: Competition and quality. *Health Affairs, 7,* 150-161.

Callen, J., Braithwaite, J., & Westbrook, J. (2006, August 20-22). In J. Westbrook & J. Callen (Eds.), *HIC 2006 bridging the digital divide: Clinician, consumer and computer Does the clinical environment affect doctors' and nurses' satisfaction with, attitudes to, and use of a computerised physician order entry system?*

Callen, J., Westbrook, J., & Braithwaite, J. (2006). The effect of physicians' long-term use of CPOE on their test management work practices. *Journal of the American Medical Informatics Association, 13,* 643-652.

Campbell, E. M., Sittig, D. F., Ash, J. S., Guappone, K. P., & Dykstra, R. H. (2006). Types of unintended consequences related to computerized provider order entry. *Journal of the American Medical Informatics Association*, *13*(5), 547-556. 10.1197/jamia.M2042.

Chaudhry, B., Wang, J., Wu, S., Maglione, M., Mojica, W., Roth, E., et al. (2006). Systematic review: Impact of health information technology on quality, efficiency, and costs of medical care. *Annals of Internal Medicine, 144*(10), 742-752.

Classen, D. C., Avery, A. J., & Bates, D. W. (2007). Evaluation and certification of computerized provider order entry systems. *Journal of the American Medical Informatics Association, 14*(1), 48-55. DOI 10.1197/jamia.M2248.

Davidson, E. J. (2005). *Evaluation methodology basics.* Thousand Oaks, CA: Sage Publications.

Davidson, E. J., & Chismar, W. (1999a). *Examining the organizational implications of IT use in hospital-based health care: A case study of computerized order entry.* Paper presented at the Proceedings of the 32nd Hawaii International Conference on System Sciences, Hawaii.

Davidson, E. J., & Chismar, W. G. (1999b). Planning and managing computerized order entry: A case study of IT-enabled organizational transformation. *Top Health Inf Manage, 19*(4), 47-61.

Department of Health. (2004). *Modernising pathology services*. Retrieved February 2, 2008, from www.dh.gov.uk/pathologymodernisation

Donabedian, A. (1988). The quality of care. How can it be assessed? *Journal of the American Medical Association, 260,* 1743-1748.

Doolan, D. F., & Bates, D. W. (2002). Computerized physician order entry systems in hospitals: mandates and incentives. *Health Affairs (Millwood), 21*(4), 180-188.

Dykstra, R. (2002). Computerized physician order entry and communication: Reciprocal impacts. In *AMIA Annual Symposium Proceedings* (pp. 230-234).

Friedman, B. A. (1996). Pathology informatics - ensuring a role as a “bit” player in laboratory medicine. *American Journal of Clinical Pathology, 105*(Suppl. 4, 1), S40-47.

Friedman, C. P., & Wyatt, J. C. (1997). *Evaluation methods in medical informatics.* New York: Springer-Verlag New York Inc.

Gell, G. (2001). Side effects and responsibility of medical informatics. *International Journal of Medical Informatics, 64*(2-3), 69-81.

Georgiou, A., & Westbrook, J. (2006). Computerised order entry systems and pathology services - a synthesis of the evidence. *Clinical Biochemist Reviews, 27*(2), 79-87.

Georgiou, A., Westbrook, J., Braithwaite, J., & Iedema, R. (2005). Multiple perspectives on the impact of electronic ordering on hospital organisational and communication processes. *HIM J, 34*(4), 130-134.

Georgiou, A., Westbrook, J., Braithwaite, J., Iedema, R., Dimos, A., & Germanos, T. (2005). A context-mechanism-outcome approach to the evaluation of computerised physician order entry systems. In H. W. Grain & M. S. Chu (Ed.), *Conference proceedings of the 13th National Health Informatics Conference.* Melbourne, Australia: Health Informatics Society of Australia.

Georgiou, A., Westbrook, J., Braithwaite, J., Iedema, R., Ray, S., Forsyth, R., et al. (2007). When requests become orders - a formative investigation into the impact of a computerized physician order entry system on a pathology laboratory service. *International Journal of Medical Informatics*, *76*(8), 583-591.

Georgiou, A., Williamson, M., Westbrook, J., & Ray, S. (2007). The impact of computerised physician order entry systems on pathology services: A systematic review. *International Journal of Medical Informatics, 76*(7), 514-529.

Greatbatch, D., Murphy, E., & Dingwall, R. (2001). Evaluating medical information systems: Etehno-methodological and interactionist approaches. *Health Services Management Research, 14*(3), 181-191.

Hardwick, D. F. (1998). Directoring and managing in a professional system. *Modern Pathology, 11*(6), 585-592.

Howanitz, J. H., & Howanitz, P. J. (2001). Laboratory results. Timeliness as a quality attribute and strategy. *American Journal of Clinical Pathology, 116*(3), 311-315.

Humber, M. (2004). National programme for information technology. *British Medical Journal, 328*(7449), 1145-1146.

Hwang, J. I., Park, H. A., & Bakken, S. (2002). Impact of a physician's order entry (POE) system on physicians' ordering patterns and patient length of stay. *International Journal of Medical Informatics, 65*(3), 213-223.

Kaplan, B. (2001). Evaluating informatics applications--some alternative approaches: Theory, social interactionism, and call for methodological pluralism. *International Journal of Medical Informatics, 64*(1), 39-56.

Kilpatrick, E. S., & Holding, S. (2001). Use of computer terminals on wards to access emergency test results: A retrospective audit. *British Medical Journal, 322*(7294), 1101-1103.

Koppel, R., Metlay, J. P., Cohen, A., Abaluck, B., Localio, A. R., Kimmel, S. E., et al. (2005). Role of computerized physician order entry systems in facilitating medication errors. *Journal of the American Medical Association, 293*(10), 1197-1203.

Kuperman, G. J., Teich, J. M., Tanasijevic, M. J., Ma'Luf, N., Rittenberg, E., Jha, A., et al. (1999). Improving response to critical laboratory results with automation: results of a randomized controlled trial. *Journal of the American Medical Informatics Association, 6*(6), 512-522.

The Leapfrog Group for Patient Safety. (2003). Fact sheet: Computerized physician order entry (CPOE). Retrieved February 2, 2008, from http://www.leapfroggroup.org/media/file/Leapfrog-Computer_Physician_Order_Entry_Fact_Sheet.pdf

McGrath, K. (2003). Review of public pathology services in Western Australia. Report to the State Health Management Team & Health Reform Committee. Retrieved February 2, 2008, from http://www.health.wa.gov.au/hrc/reports/docs/Pathology_Services_Review.pdf

Mekhjian, H. S., Kumar, R. R., Kuehn, L., Bentley, T. D., Teater, P., Thomas, A., et al. (2002). Immediate benefits realized following implementation of physician order entry at an academic medical center. *Journal of the American Medical Informatics Association, 9*(5), 529-539.

Murff, H. J., Gandhi, T. K., Karson, A. K., Mort, E. A., Poon, E. G., Wang, S. J., et al. (2003). Primary care physician attitudes concerning follow-up of abnormal test results and ambulatory decision support systems. *International Journal of Medical Informatics, 71*(2-3), 137-149.

National Electronic Decision Support Taskforce. (2003). *Electronic decision support for Australia's health sector.* Canberra, Australia: Commonwealth of Australia.

Neilson, E. G., Johnson, K. B., Rosenbloom, S. T., Dupont, W. D., Talbert, D., Giuse, D. A., et al. (2004). The impact of peer management on test-ordering behavior. *Annals of Internal Medicine, 141,* 196-204.

Nevalainen, D., Berte, L., Kraft, C., Leigh, E., Picaso, L., & Morgan, T. (2000). Evaluating laboratory performance on quality indicators with the six sigma scale. *Archives of Pathology & Laboratory Medicine, 124*(4), 516-519.

Nightingale, P. G., Peters, M., Mutimer, D., & Neuberger, J. M. (1994). Effects of a computerised protocol management system on ordering of clinical tests. *Quality Health Care*, *3*(1), 23-28.

NSW Government Action Plan for Health. (2002). *Information management and technology education, training and development strategy.* Retrieved February 2, 2008, from http://www.health.nsw.gov.au/policy/gap/strategies/im&t_strategy.pdf

Ostbye, T., Moen, A., Erikssen, G., & Hurlen, P. (1997). Introducing a module for laboratory test order entry and reporting of results at a hospital ward: An evaluation study using a multi-method approach. *Journal of Medical Systems, 21*(2), 107-117.

Overhage, J. M., Tierney, W. M., Zhou, X.-H., & McDonald, C. J. (1997). A randomized trial of corollary orders to prevent errors of omission. *Journal of the American Medical Informatics Association, 4*(5), 364-375.

Pawson, R. (2004). Evidence-based policy: A realist perspective. In B. Carter & C. New (Eds.), *Making realism work; realist social theory and empirical research* (pp. 26-49). London: Routledge.

Plebani, M., & Carraro, P. (1997). Mistakes in a stat laboratory: Types and frequency. *Clinical Chemistry, 43*(8 Pt 1), 1348-1351.

Poon, E. G., Gandhi, T. K., Sequist, T. D., Murff, H. J., Karson, A. S., Bates, D. W., et al. (2004). I wish I had seen this test result earlier!: Dissatisfaction with test result management systems in primary care. *Archives of Internal Medicine, 164*(20), 2223-2228.

Potter, J. (2000). Standards of care for older people: An introduction to the issues. In J. Potter, A. Georgiou & M. Pearson (Eds.), *Measuring the quality of care for older people*. London: Royal College of Physicians.

Review of NHS Pathology Services in England. (2006). Report of the Review of NHS Pathology Services in England. Retrieved February 2, 2008, from www.dh.gov.uk

The Royal College of Pathologists. (2000). Pathology: The hidden science that saves lives. Retrieved February 2, 2008, from http://www.rcpath.org/index.asp?PageID=603

Scriven, M. (1991). *Evaluation thesaurus* (4th ed.). Newbury, CA: Sage.

Shekelle, P., Morton, S., & Keeler, E. (2006, April). *Costs and benefits of health information technology* (Evidence report/technology assessment No. 132 AHRQ publication No. 06-E006). Rockville, MD: Agency for Healthcare Research and Quality.

Shu, K., Boyle, D., Spurr, C., Horsky, J., Heiman, H., O'Connor, P., et al. (2001). Comparison of time spent writing orders on paper with computerized physician order entry. MEDINFO 2001. In *Proceedings of the 10th World Congress on Medical Informatics* (pp. 1207-1211).

Sittig, D., & Stead, W. (1994). Computer-based physician order entry: The state of the art. *Journal of the American Medical Informatics Association, 1*(2), 108-123.

Smith, B. J., & McNeely, M. D. (1999). The influence of an expert system for test ordering and interpretation on laboratory investigations. *Clinical Chemistry, 45*(8 Pt 1), 1168-1175.

Solomon, D. H., Shmerling, R. H., Schur, P. H., Lew, R., Fiskio, J., & Bates, D. W. (1999). A computer based intervention to reduce unnecessary serologic testing. *Journal of Rheumatology, 26*(12), 2578-2584.

Stablein, D., Welebob, E., Johnson, E., Metzger, J., Burgess, R., & Classen, D. C. (2003). Understanding hospital readiness for computerized physician order entry. *Joint Commission Journal on Quality Safety, 29*(7), 336-344.

Stoop, A. P., & Berg, M. (2003). Integrating quantitative and qualitative methods in patient care information system evaluation: guidance for the organizational decision maker. *Methods of Information in Medicine, 42*(4), 458-462.

Tetrault, G. A. (2001). Clinical laboratory quality assurance. In J. B. E. Henry (Ed.), *Clinical diagnosis and management by laboratory methods* (pp. 148-158). Philadelphia, PA: WB Saunders Company.

Tierney, W. M., Miller, M. E., & McDonald, C. J. (1990). The effect on test ordering of informing physicians of the charges for outpatient diagnostic tests. *New England Journal of Medicine, 322*(21), 1499-1504.

Travers, E. (1997). Management of information and its economic impact. In E. Travers (Ed.), *Clinical Laboratory Management.* Baltimore, MD: Williams & Wilkins.

van Walraven, C., & Naylor, C. D. (1998). Do we know what inappropriate laboratory utilization is? A systematic review of laboratory clinical audits. *Journal of the American Medical Association, 280*(6), 550-558.

Wears, R. L., & Berg, M. (2005). Computer technology and clinical work; still waiting for Godot. *Journal of the American Medical Association, 293*(10), 1261-1263.

Westbrook, J. I., Braithwaite, J., Iedema, R., & Coiera, E. W. (2004). Evaluating the impact of information communication technologies on complex organizational systems: A multi-disciplinary, multi-method framework. *Medinfo, 11*(Pt 2), 1323-1327.

Westbrook, J., Georgiou, A., Dimos, A., & Germanos, T. (2006). Computerised pathology test order-entry reduces laboratory turnaround times and influences test ordered by hospital clinicians: A controlled before and after study. *Journal of Clinical Pathology, 59,* 533-536.

KEY TERMS

Electronic Decision Support Systems: Access to knowledge stored electronically to aid patients, carers, and service providers in making health care decisions (National Electronic Decision Support Taskforce, 2003).

Computerised Provider Order Entry: Electronic systems that allow physicians (or other authorised staff) to enter hospital orders directly into a computer.

Evaluation: To determine the merit, worth, or value of something, or the product of that process (Scriven, 1991).

Impact: Change or (sometimes) lack of change caused by an evaluand (that which is being evaluated). Similar in meaning to outcome and effect (Davidson, 2005).

Indicator: A factor, variable, or observation that is empirically connected with the criterion variable (Scriven, 1991).

Pathology: Clinically-led diagnostic, laboratory, and post mortem services that cover a range of tests on blood and other human materials necessary for diagnosis and monitoring of a wide range of clinical conditions so that appropriate treatment can be given (Department of Health, 2004).

Quality: Merit or the extent to which an evaluand meets identified needs and relevant standards (Davidson, 2005).

Redundant tests: A reordered or repeated laboratory test that is ordered within an inappropriate time frame and provides no additional information (Bates et al., 1998; van Walraven & Naylor, 1998).

Turnaround Times (TAT): A frequently used measure by pathology services. Total TAT can be defined as the time of physician order request to when the physician reviews the result. Laboratory TAT measures the time a specimen arrives at the laboratory to the time of results dispatch.

Emerging Technologies for Aging in Place

Shirley Ann Becker
Florida Institute of Technology, USA

Frank Webbe
Florida Institute of Technology, USA

INTRODUCTION

Extraordinary societal changes in the United States have taken place from both human and technological perspectives. Life expectancy continues to increase, due to improvements in both medical care and technology. Older adults, 65 years plus, comprise about 12.4% of the United States population with about one in every eight Americans being in this age group (U.S. Administration on Aging, 2002). By 2030, the percentage will have increased to 20% of the total population representing twice the number as in 2000. Those 85 years and older represent the fastest growing group (He, Sengupta, Velkoff, & DeBarros, 2005).

Older adults in the United States not only are living longer, but many are staying longer in their homes before entering institutionalized care. Compared to past generations, an unprecedented number of adults sixty years and older are aging in place. In 1985, there were approximately 5.5 million functionally disabled elderly living in their communities and an additional 1.3 million living in nursing homes. Each of these figures is projected to double to 10.1 million and 2.5 million; respectively, by 2020 (Manton, 1989). This is significant from both individual and national perspectives in terms of the burden of skyrocketing healthcare costs.

The U.S. Bureau of the Census (Fields & Casper, 2001) has reported that more than 55% of older adults live at home with their spouse. The number of adults living *without* a spouse increases with age, with about 50% of women aged 75 years and older living alone at home. Many older adults prefer to age in place rather than in an assisted living facility (Riche & MacKay, 2005), citing independence and social interaction as being critical to their well-being (Hirsch, Forlizzi, Hyder, Goetz, Kurtz, & Stroback, 2000). Complicating the choice to age in place are health and other personal limitations that necessitate continuous care. About seven million older adults over 65 years are estimated to have mobility or self-care limitations (U.S. Administration on Aging, 2002).

Similar to other nations, the United States faces a critical challenge in dealing with an aging population that has unprecedented life expectancies. Emerging technologies offer the hope of allowing older adults to remain in their homes longer by empowering individuals to manage daily activities while dealing with chronic health conditions and age-related diseases. These technologies increasingly target a home environment whereby on a regular basis an individual can obtain assistance in performing daily living activities, stay connected to family and friends, manage medication, and be monitored for health-related changes. As important as these assistive technologies are for individuals and families, their potential for positively impacting the United States economy by changing the model of healthcare delivery is equally huge.

BACKGROUND

Recent advances in "assistive technologies" offer older adults the ability to function independently and to age with dignity in their homes and communities. The U.S. Administration on Aging (2005) formally defines an assistive technology as being any service or tool that helps the elderly or disabled perform activities they have always done, but must now do differently. Assistive technology includes communication equipment, computer access, tools for independent living, education, and mobility aids, among others. Table 1 provides a summary of current and emerging assistive technologies.

The availability of assistive technologies for an older adult often determines whether he or she is able to live independently or must move to an institutionalized environment. The National Council on Disability (1993) found that 80% of older adults who used assis-

Table 1. Types of assistive technologies with emerging trends

Type	Assistive Technology
Adaptive Switches	Devices that are used to adjust home systems, including air conditioners, computers, answering machines, and power wheelchairs among others. Emerging technologies include "smart homes" with sensing systems that manage the overall home environment.
Communication Aids	Technology that promotes social interaction, telecare, and other types of communication to promote functional independence. Emerging technologies include the use of Personal Digital Assistants (PDAs), personal computers, and tablets to connect an older adult to an external support system of family, friends, or healthcare personnel.
Computer Accessibility Devices	Software or hardware that supports computer and Web accessibility. This includes screen reader devices, specially designed keyboards, and others. Emerging technologies include user-interface designs of computer, handheld, and mobile devices that promote universal access.
Home Modifications	Construction or modifications that allow an older adult to overcome a physical barrier that might impede living at home. Emerging technologies are illustrated by sensor devices in the floors to detect falls.
Independent Living Aids	Technology that allows an older adult to function independently in the home. Grab bars in bathrooms and hallways are more traditional illustrations of this type of technology. Emerging technologies include medication management (e.g., talking medicine cabinet), and electronic prompts (e.g., close refrigerator), among others.
Mobility Aids	Technology that supports an older adult being mobile within and external to the home environment. Emerging technologies include Global Positioning Systems (GPS) built into mobile devices as a means of identifying a person's location outside the home.

tive technology were able to reduce their dependence on others. Assistive technologies not only support the aging adult, but also the family and friend caregivers. Most often, technologies that increase the independence of an older adult will decrease the time required for caregiving assistance (Mann, 2001). Assistive technology and home modifications have also been found to provide caregivers immediate relief from burden, reduce their stress, and help them provide care more easily and safely (Gitlin, Corcoran, Winter, Boyce, & Hauck, 2001). This is significant from an aging in place perspective, given that many adults with chronic conditions and age related diseases are cared for by an aging spouse may also suffer from a chronic medical condition.

Many of the emerging technologies are being developed to be used in both home and community environments. They provide an opportunity to improve independence, and safety thus promoting aging in place initiatives (Tran, 2004). Some research efforts are building on existing technologies in order to take advantage of communication and information infrastructures. The Internet along with online information sources, for example, offer a means for older adults to engage in lifelong learning and healthcare management thus promoting quality of life while aging in place. Recent advances in wireless and wired communication provide the technology for home monitoring systems, social interaction, and the potential for virtual support systems linking an older adult to geographically dispersed family and friends. Emerging areas of technology include artificial intelligence and robotics with the opportunity of providing intelligent feedback to the elderly living at home.

EMERGING TECHNOLOGIES

Two types of technologies are being developed currently that promote aging in place depending on the *category of impairment* associated with the individual. (Refer to the definitions in Appendix A of this article for a description of the three levels of impairment). Becker and Webbe's (2006b) Buddy Coordinated Healthcare System, and Scott and Gabrielli's (2005) Ho'alauna ("Good Neighbor") Tablet projects focus on older adults who have mild or moderate levels of impairment. The underlying concept is to utilize technology to promote independent functioning in both home and community environments. These projects are considered "noninvasive" in that the individual has control over data gathering and dissemination using the proposed technologies (Becker & Webbe, 2006a).

Other research projects, such as the Digital Family Portrait (Mynatt, Rowan, Craighill, & Jacobs, 2001) and the CareNet Display (Consolvo, Roessler, & Shelton, 2004; Consolvo & Towle, 2005) would support more severe levels of impairment through a home monitoring environment that utilizes sensors to gather information about daily living activities. Such detection provides the means to keep members of a support network (e.g., family, friends, and healthcare personnel) informed of the older adult's daily activities. These types of "invasive" technologies do not provide the older adult full control over data gathering and dissemination activities. These are automated as part of the home monitoring system that is put in place. (Table 2 further describes the aforementioned noninvasive and invasive technology research projects as illustrative of emerging technologies.)

These are only several of a significant number of research projects currently being conducted in university and laboratory settings. Many related projects have been or are being developed to promote aging in place. Academia and industry partners have answered the call to innovate technologies to address the looming healthcare crisis, due to the aging baby boomer population in the United States. One of the biggest obstacles, however, is that many emerging technologies are not being translated into products and services that are made available to older adult users. Eric Dishman of Intel Corporation (Beck, 2004) told the Senate Special Committee on Aging, "Our biggest problem nationally is an imagination problem, not a technology problem," in terms of government sponsoring research on assistive technologies for older adults with corporate investment opportunities.

Table 2. Assistive technologies illustrating the use of information and communication technologies (ICT) to promote aging in place

Project	Description
Buddy Coordinated Healthcare System	Handheld technology is used by the aging adult caregiver to assist in caregiving activities and monitor his or her own well-being and that of the older adult for whom care is provided. A PocketPC is used to gather information about daily living activities, emotions of the caregiver, behaviors displayed by the person receiving care, medication management, and other support systems. The PocketPC has the capability to transparently and unobtrusively link to the Internet thus sharing information gathered on the device via a family and friend web site. Family and friends use the Web site to foster sharing of the responsibilities associated with caregiving. Family and friends may otherwise not be actively involved due to geographic distance, work, children, and other commitments (Becker & Webbe, 2006b).
Digital Family Portrait	This project is part of Georgia Tech's Aware Home Research Initiative (http://www.awarehome.gatech.edu/). It provides a virtual connection between geographically distant family members to an older adult living at home. The technology, when fully developed, utilizes sensors to gather data about the older adult's health, relationships, and activities. This information would be shared electronically via the ambient display placed in a distant relative's home. The ambient display, presented as a digital picture frame, provides user access to detailed information. The information presented to the user is similar to information typically observed by a neighbor or another member living in the older adult's household (Mynatt et al., 2001).
Ho'alauna Tablet	The Ho'alauna Tablet project utilizes ICT to assist with nutrition, healthcare, financial management, and leisure activities, among others (Scott & Gabrielli, 2004). It is proposed that a TabletPC along with specialized software supported by Internet access assist older adults living at home to interact socially with family and friends, search the web, control home technology, receive emergency information and assistance, track health-related information, and participate in community events, social activities, and educational programs.
CareNet Display	This project uses the concepts of ambient display and home sensors to build a support network of family and friends (Consolvo et al., 2004). The ambient display, an interactive digital picture frame, provides information to a support network regarding the status of meals, medications, outings, activities, mood, falls, and schedules. The ambient display allows the user to view an older adult's daily activities and interactively access data through touch screen capability. Unlike the Digital Family Portrait, it allows for the sharing of sensitive data regarding medication and activities performed. The older adult would control shared access to personal information. The CareNet Display allows for data sharing of other members in the support network.

We have identified four underlying issues that are directly related to the crisis associated with technology transfer from the academic research setting to those who could benefit from these technologies. These issues have the potential to pose barriers to continued research on emerging technologies, thus impeding the efforts to promote aging in place. As such, these issues must be addressed to facilitate the development of assistive technologies. The issues and directions of development are described briefly:

1. **Usability in a real world setting:** Researchers need access to older adult volunteers to identify potential usability barriers that would impede technology transfer. Although the new technologies can be prototyped with young volunteer users, the special demands of the older end-user require in-vivo testing and iterative hardware and software modifications before introduction into a home environment.
2. **Patient safety:** Research that introduces new assistive technologies into the home environment must complement the human caregivers and take into account an increase in safety concerns. For example, existing caregivers may assume that the new technology and its keepers "looks out" for them and their family member so that normal precautions are reduced, including contact with their medical professionals. It is incumbent upon the researchers to train the caregivers that the new technology is truly *assistive* to them and under their control.
3. **Interoperability:** Currently, many projects are developed in isolation of one another, even though they may ultimately coexist within common venues. There must be a substantially greater national push for interoperability of technologies, even at the university level during the developmental phases. The field of assistive technologies is too fragile at the moment to host a "CDMA vs. GSM" type of competition.
4. **Social perceptions:** Corporations who shy away from the aging marketplace must be convinced that the potential is significant. The government has to be innovative in providing research funding not only to innovate but to transfer the technology to the commercial sector.

CONCLUSION

The United States and other countries are facing a significant challenge in dealing with the exploding healthcare costs associated with an aging population. There are already 34 million senior citizens in the United States with an impending explosion as the baby boomers begin to enter the retirement phase of their lives. There has to be an increased effort to provide both invasive and noninvasive assistive technologies to manage proactively the health and well-being of the aging population. This includes a significant emphasis on aging in place in order to reduce the costs associated with assisted living facilities and hospitals when older adults no longer can remain in their homes and communities.

Emerging technologies for aging in place provide additional benefits to a younger population with disabilities and chronic diseases. For many middle aged adults with chronic health conditions, assistive technologies provide the means to monitor healthcare proactively. For diseases such as diabetes, hypertension, arthritis, and others, assistive technologies offer the hope of better living from both longevity and quality of life perspectives. It would be important to address the issues of usability, safety, interoperability, and social perceptions as an integral part of assistive technology research.

Researchers, nonprofit organizations, industry, and the government are initiating collaboration in order to address these issues. Although more needs to be done to promote this collective effort to innovate technologies for aging in place, two recent initiatives serve as prototypes of the productive coalitions can be formed. The American Association of Homes and Services for the Aging (http://www.AAHSA.org) and Intel Corporation, for example, initiated the launching of the Center for Aging Services Technologies (CAST) with the objective of advancing technologies for older adults. The Alzheimer's Association along with industry partners has developed the Everyday Technologies for Ahzheimer's Care (ETAC) consortium to bring university researchers, nonprofit organizations, government, and industry together to advance assistive technology development while addressing the underlying issues identified in this article. These models both show the feasibility of academic/industrial/agency coalitions, and also through their already considerable outcomes show what products can ensue.

E

REFERENCES

Beck, E. (2004, April 27). Seniors need robots and new technology to help at home. [Electronic version]. *SpaceDaily*.

Becker, S. A., & Webbe, F. M. (2006a). Designing for older adult users of handheld technology. In *Proceedings of the 28th IEEE Engineering in Medicine and Biology Society Annual International Conference* (pp. 3297–3300).

Becker, S. A., & Webbe, F. (2006b). Use of handheld technology by older adult caregivers as part of a virtual support network. In *Proceedings of the First International Conference on Pervasive Computing Technologies for Healthcare 2006.* Innsbruck, Austria.

Consolvo, S., Roessler, P., & Shelton, B. E. (2004). The CareNet display: Lessons learned from an in home evaluation of an ambient display. In *Proceedings of the 6th International Conference on Ubiquitous Computing: UbiComp'04* (pp. 1–17). Nottingham, England.

Consolvo, S., & Towle, J. (2005). Evaluating an ambient display for the home. In *Proceedings of the Computer Human Interaction CHI'05 Conference*. Portland, OR.

Fields, J., & Casper, L. M. (2001). *America's families and living arrangement population characteristics. P20-537*. U.S. Bureau of the Census, U.S. Department of Commerce, Washington, D.C.

Gitlin, L., Corcoran, M., Winter, L., Boyce, A., & Hauck, W. (2001). A randomized controlled trial of a home environmental intervention: Effect on efficacy and upset in caregivers and on a daily function of persons with dementia. *Gerontologist 41*(1), 4–14.

He, W., Sengupta, M., Velkoff, V. A., & DeBarros, K. A. (2005). *65+ in the United States: 2005.* Special report issued by the U.S. Department of Health and Human Services and the U.S. Department of Commerce, Washington, DC.

Hirsch, T., Forlizzi, J., Hyder, E., Goetz, J., Kurtz, C., & Stroback, J. (2000). The ELDer project: Social, emotional, and environmental factors in the design of eldercare technologies. In *Proceedings of CM Conference on Universal Usability* (pp. 72–79). Arlington, VA.

Kohn, L. T., Corrigan, J. M., & Donaldson, M. S. (2000). *To err is human: Building a safer health system.* A report of the Committee on Quality of Health Care in America, Institute of Medicine. Washington, DC: National Academy Press.

Mann, W. C. (2001). The potential of technology to ease the care provider's burden, *Generations*, *25*(1), 44–49.

Manton, K. G. (1989). Epidemiological, demographic and social correlates of disability among the elderly. *The Milbank Quarterly*, *67*(2), 13–58.

Mynatt, E. D., Rowan, J., Craighill, S., & Jacobs, S. (2001). Digital family portraits: Supporting peace of mind for extended family members. In *Proceedings of Computer Human Interaction CHI'01 Conference* (pp. 333–340). Seattle, WA.

National Council on Disability. (1993, March). *Study on the financing of assistive technology devices and services for individuals with disabilities: A report to the President and the Congress of the United States*. Retrieved February 14, 2008, from NCD Newsroom http://www.ncd.gov/newsroom/publications/1993/assistive.htm.

Riche, Y., & Mackay, W. (2005). Peercare: Challenging the monitoring approach to care for the elderly. In *Proceedings of the British Human Computer Interaction 2005 Workshop on HCI and the Older Population.* Edinburgh, United Kingdom.

Scott, N., & Gabrielli, S. (2004). *Overview of the Ho'alauna tablet.* Archimedes Project, University of Hawaii. Retrieved February 14, 2008, from http://archimedes.hawaii.edu/Downloads/HOALAUNA2.pdf.

Tran, B. Q. (2004). Technologies to facilitate health and independent living in elderly populations. In D.C. Burdick, & S. Kwon (Eds.), *Gerotechnology Research and Practice in Technology and Aging* (pp. 161–173), New York, NY: Springer.

U.S. Administration on Aging. (2002). *A profile of older Americans.* Administration on Aging, U.S. Department of Health and Human Services. Retrieved February 14, 2008, from http://www.aoa.gov/prof/Statistics/profile/2002profile.pdf

U.S. Administration on Aging. (2005). *What is assistive technology fact sheet.* U.S. Department of Health and

Human Services. Retrieved February 14, 2008, from http://www.eldercare.gov/eldercare/Public/resources/fact_sheets/assistive_tech_pf.asp

KEY TERMS

Aging in Place: Older adults remain in their homes and communities through the support of technology and assisted living services.

Assistive Technology: A device or system supporting an individual in performing a task that otherwise could not be done, or increases the ease and safety with which it is done.

Impairment Categories: The U.S. Department of Health and Human Services (http://www.hhs.gov) provides three categories of impairments with varying assistive technology needs. The *Mildly Impaired* category (e.g., mild arthritis) benefits from assistive technologies that enhance independent living. This may include special devices such as grab bars and modified cooking utensils. The second category, *Moderately Impaired,* include those who are functionally impaired in several daily living activities. For example, a person with arthritis may also have diabetes with circulation problems. Assistive technology supplements other types of support systems (e.g., home caregiver) in order to promote independent living. The third category, *Severely Impaired,* probably will benefit the most from emerging technologies. These individuals have multiple impairments that impact daily living.

Healthcare Technology: This term is typically used to encompass all technologies that are used in the healthcare field inclusive of both medical personnel and patients.

Patient Safety: Institute of Medicine defines patient safety as "freedom from accidental injury;" conversely, error constitutes "the failure of a planned action to be completed as intended or the use of a wrong plan to achieve an aim." (Kohn, Corrigan, & Donaldson, 2000).

Proactive Computing: Applications that are designed to anticipate an individual's needs and to take action to meet the needs on their behalf (taken from http://www.Intel.com's Web site on Health Research and Innovation).

Smart Home Technology: Devices used in a home environment to perform complex tasks that otherwise would be performed by the individual. They facilitate independent living by providing automated support of daily activities, security, building performance, and telecommunications.

Technology Assistance Act: Congress enacted the Technology Related Assistance Act of 1988 (P.L.100-407) to expand the availability of assistive technology services and devices to people with disabilities.

Enhanced Rheoencephalography

Juan J. Perez
Polytechnic University of Valencia, Spain

Enrique Guijarro
Polytechnic University of Valencia, Spain

Pedro Ortiz
Consorcio Hospital General Universitario de Valencia, Spain

José M. Pons
Consorcio Hospital General Universitario de Valencia, Spain

INTRODUCTION

A close monitoring of the cerebrovascular parameters is essential in some neurological diseases to prevent secondary brain insults. Persistent rises in intracranial pressure caused by oedema, tumours, or haematomas may decrease the cerebral blood flow (CBF) to values below the minimum required for neuronal survival (Lang & Chesnut, 1994).

As a method to evaluate CBF, Polzer and Schuhfried proposed in 1950 to apply the well-known impedance plethysmography techniques to the head, which was specifically referred to as Rheoencephalography (REG) (Polzer & Schuhfried, 1950). For this purpose, bipolar and tetrapolar electrode arrangements, named as REG I and REG II respectively, were used to measure the impedance changes of the head synchronized to the heartbeat. In REG I, a low-amplitude current is injected through two electrodes attached to the scalp, and the related electric potential difference is measured between these electrodes, whereas in REG II, a second pair of electrodes is used to measure the electric potential difference (Geddes & Baker, 1989). The latter technique reduces eventual artifacts caused by mechanical alterations in the electrode-skin interface of the current injection electrodes.

The REG signal was assumed to be induced by the time-varying nature of the mixture ratio of two biological conductors with different electrical conductivity: namely, blood and brain. According to this, REG would reflect the changes in cerebral blood volume (CBV) associated with the cardiac cycle (Lifshitz, 1970).

However, despite the great efforts made by the scientific community on REG research during three decades, the use of REG was not extended to the clinical practice after all, because REG ability to reflect the CBF was strongly questioned (Basano, Ottonello, Nobili, Vitali, Pallavicini, Ricca, Prastaro, Robert, & Rodríguez, 2001; Perez-Borja & Meyer, 1964). Detractors argued that a significant part of the injected current does not cross the skull, since the electrical conductivity of the skull is approximately 80 times lower than that of the scalp and brain. This means that most of the REG signal reflects the pulsatile changes in the scalp blood volume (SBV), rather than in the CBV (Laitinen, 1968). This disagreement, together with the development of the transcranial Doppler ultrasound technology, led to the abandonment of research on REG.

BACKGROUND

A sound review of the literature reveals that research on REG was very intensive during the fifties and sixties, due to the lack of reliable diagnosis tools for the assessment of cerebral blood flow (Hadjiev, 1972). However, modern neuroimaging techniques provide a great amount of information about the brain metabolism for diagnosis purposes. Yet, the idea of a noninvasive, portable, available-at-bedside, and low-cost device for real-time monitoring of the cerebrovascular system still seems attractive and necessary.

Recently, the physical principles of REG have been analyzed from a theoretical perspective for a better understanding of the mechanisms, by which the extra and intracranial blood flows contribute to the REG signal (Perez, Guijarro, & Barcia, 1999; Perez, Guijarro, & Barcia, 2000; Perez, Guijarro, & Barcia, 2004). The main findings of these theoretical studies explain the initial controversy on the origin of the REG signals, and can be summarized as follows: (i) most of the REG I signal is caused by the pulsatility of the extracranial blood flow; (ii) in some subjects, a set of tetrapolar electrode arrangements may exist to record a REG II free of extracranial information; and (iii) such electrode positions, and even their existence, are strongly dependent on the subject's physical constitution, particularly on his/her scalp thickness. In summary, previous theoretical results suggest that there is no universal electrode arrangement suitable for all individuals to register a REG II free of extracranial contamination.

Therefore, a REG II recorded from an arbitrary tetrapolar electrode arrangement contains information from both intra and extracranial blood flows, mixed in unknown proportions that, furthermore, depend on the used electrode arrangement. Attempts have been made in the past to cancel out the extracranial component from the REG signal. For instance, Seipel (1967) proposed a method to identify the contribution to REG by each one of the intra and extracranial arteries that supply the brain and scalp. It consisted in occluding some of the arteries, and then deducing their contribution by means of simple equations. However, this method was experimentally tested by Masucci, Seipel, and Kurtzke (1970), who did find no value for detecting, lateralizing, or diagnosing cerebral injury.

One unexplored way of extracting intracranial information from a REG signal could be to address the problem using the Blind Source Separation (BSS) approach. In BSS, a set of unknown signal sources is guessed from the analysis of a set of observations, each one of which is a mixture in unknown proportions of the sources. The extraction of the sources from the observations is achieved by taking advantage of some assumed property of the sources (Cardoso, 1998).

A practical example of BSS is the so-called cocktail party problem, in which the statistical independence of the sources is assumed, and its solution is reached by using the well-established Independent Component Analysis (ICA). Let us imagine a room where, for instance, three people are talking simultaneously. Three microphones placed in the same room at different sites will record a mixture of the speakers' voices, whose weights will depend on the speakers' distance to the microphones. This information mixture can be expressed as:

$$x_1(t) = a_{11}s_1(t) + a_{12}s_2(t) + a_{13}s_3(t) \quad (1)$$

$$x_2(t) = a_{21}s_1(t) + a_{22}s_2(t) + a_{23}s_3(t) \quad (2)$$

$$x_3(t) = a_{31}s_1(t) + a_{32}s_2(t) + a_{33}s_3(t) \quad (3)$$

where $s_i(t)$ are the speakers' voice signals, $x_i(t)$ the observations, and a_{ij} the mixture coefficients (i, j = 1,.., 3). Observe that both the mixture coefficients and the sources are unknown in these equations. In this example, it is reasonable to assume the statistical independence of the sources, so ICA can be applied to guess the sources and coefficients. An excellent illustration of the cocktail party problem, and ICA, can be found in Hyvärinen, Karhunen, and Oja (2001).

This mathematical tool could be of great interest in REG since, as commented above, a REG II signal can be regarded as the weighted sum of intra and extracranial information. However, statistical independence cannot be used here to separate the components, since scalp and brain perfusions are caused by the same event: the heart contraction. Therefore, to apply the BSS approach to REG, it is necessary to find a differential property between both components that could be mathematically formulated.

EXTRACTION OF THE INTRACRANIAL COMPONENT FROM THE REG: ENHANCED RHEOENCEPHALOGRAPHY

Mathematical Formulation

Let $R_1(t)$ and $R_2(t)$ be a pair of REG signals recorded simultaneously from a given subject at different electrode arrangements. Assuming the morphological invariability and negligible phase shift of both REG components (Perez, Guijarro, & Sancho, 2005), we can write:

$$R_1(t) = a_{11}C_{En}(t) + a_{12}C_{In}(t) \quad (4)$$

$$R_2(t) = a_{21}C_{En}(t) + a_{22}C_{In}(t) \quad (5)$$

where $C_{En}(t)$ and $C_{In}(t)$ are, respectively, the extra and intracranial components of the recorded REG signals normalized to unit variance (sources), and a_{ij} are the mixture coefficients that depend on the electrode arrangement. Both components and coefficients of the mixture matrix are unknown. Now, by subtracting an arbitrary ratio *K* in (4) from (5), we can define a function $F_K(t)$ as:

$$F_K(t) = R_2(t) - K\,R_1(t) = \\ (a_{21} - K\,a_{11})C_{En}(t) + (a_{22} - K\,a_{12})C_{In}(t). \qquad (6)$$

Then, the goal is to find a *K** value that cancels out the extracranial component term in (6). For this purpose, we exploit the following difference between the intra and extracranial blood perfusions. On the one hand, the quick rise in arterial blood pressure that follows a heartbeat causes an abrupt increment in the extracranial blood volume, which is responsible for the sharp drop of the extracranial component (Perez et al., 2005). In fluid dynamics terms, the extracranial blood flow measured, for instance, at the superficial temporal artery, will show a markedly pulsatile waveform. On the other hand, fast increases in CBV are not allowable within the rigid cranial structure, and therefore, CBF should show, comparatively, a smoother waveform. According to this, since *K* modulates the amount of extracranial blood volume information contained in $F_K(t)$, the searched *K** value will give a minimum value of the function cost:

$$J(K) = \left[\frac{d}{dt}\left(F_K(t)\right)\right]_{RMS}$$

where $J(K)$ measures the pulsatility of the $F_k(t)$ function, and the term RMS refers to the root-mean-square value. Then, the *K** value that minimizes (7) will transform (6) into:

$$F_{K^*}(t) = R_2(t) - K^*\,R_1(t) = (a_{22} - K^*\,a_{12})C_{In}(t) \qquad (7)$$

Although this procedure could be used to extract the morphology of the intracranial component from two REG II signals, a more interesting result can be obtained by using a REG I signal as $R_1(t)$. In this case, and assuming that the REG I signal is caused exclusively by the extracranial blood flow, the a_{12} term in (1) can be considered negligible, and $F_K{}^*(t)$ becomes the intracranial component of $R_2(t)$. This assumption agrees with previous clinical (Perez-Borja & Meyer, 1964) and theoretical (Perez, Guijarro, Sancho, & Navarré, 2006) studies.

Experimental Findings

Current research on this separation method focuses on demonstrating that the extracted REG component is actually of intracranial origin. This requires several approximations because there is no any gold-standard to compare the extracted REG component with other real-time CBV measurements.

In a first stage, the morphological invariability of the extracted component can be studied: since the morphology of the intra and extracranial components are different, the morphology of REG II should be expected to depend strongly on the used electrode arrangement used, whereas the morphology of the extracted component should keep constant throughout the scalp surface. To study this influence, the following experiment was performed on volunteers.

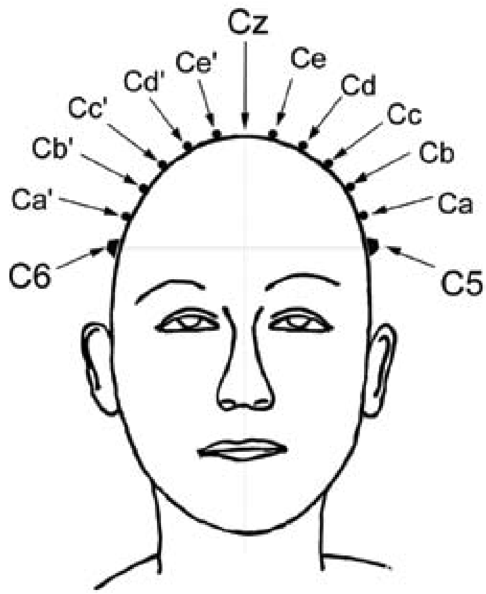
Twelve electrode sites were properly marked and cleaned on the subject's scalp surface: C5; C6; five equidistant electrode sites between C5 and Cz (labelled as Ca, Cb, Cc, Cd and Ce); and other five between C6 and Cz (labelled as Ca', Cb', Cc', Cd' and Ce') (Figure 1). The current injection electrodes were permanently fixed at C5 and C6, whereas the position of the REG II pickup electrodes was changed at each recording. This set of electrode arrangements has been chosen because it matches well with that used in previous theoretical studies (Perez et al., 2000; Perez et al., 2004). An additional pair of electrodes was used to record ECG (Lead I) for synchronization purposes.

Five records of REG I, REG II, and ECG channels were acquired during three minutes. While the current injection electrodes were kept at C5 and C6 during all the experiments, the pairs of pickup electrodes used at each recording stage were placed at sites Ca-Ca'; Cb-Cb'; Cc-Cc'; Cd-Cd'; Ce-Ce'.

To improve signal-to-noise ratio, REG sweeps time-locked with the R wave of the ECG were averaged using a time-window extended from 100 ms prior to the R wave, to 700 ms after it. The aforementioned method was used to extract the intracranial component of the REG II signal for each pair of the five averaged data sets.

Averaged REG I waveforms are shown overlapped in Figure 2a. For clarity, the traces have been normalized to unit variance; in this way, they can be read as $C_{En}(t)$ components if we assume the premise that REG I is of pure extracranial origin. The traces show a morphology very close to that previously identified

Figure 1. Schematic representation of the electrode positions used in the experimental section. Tetrapolar impedances were measured by injecting current through electrodes C5 and C6, whereas electrode pairs symmetrical to the coronal plane were used for voltage measurement



as extracranial REG component (Perez et al., 2005). The time interval in which the impedance rising edge occurs is marked on Figure 2a and transferred to the next figures for comparison. As expected, the morphology of the REG I traces is practically constant along the recordings since the injection electrodes keep invariable throughout the experiment.

On the contrary, a high variability of the $R_2(t)$ traces can be observed in Figure 2b. Statistical analysis of the correlation coefficients between each pair of $R_2(t)$ traces can be found in Perez et al. (2006). Within the time interval of the $R_1(t)$ rising edge previously presented in Figure 2a, replicas of that edge can be observed in the two electrode pairs closer to the injection one (Ca-Ca' and Cb-Cb'). Moreover, a rising edge of the impedance trace can also be observed, but unexpectedly inverted, in the impedance measurement recorded at the electrode pair placed farthest from the injection one (Ce-Ce'). This paradoxical result was predicted in Perez et al. (2004), and means that the sign of the influence (i.e., the sensitivity) of the extracranial component on REG II changes. Finally, all the intracranial $C_{In}(t)$ components extracted from R_2(t) traces show similar morphologies (figure 3). Within the R_1(t) time interval represented in Figure 2a, no sign of the rising edge is observed, the waveforms being similar to the R_2(t) traces recorded at intermediate electrode positions (Cc-Cc' and Cd-Cd').

Several facts suggest an intracranial origin of the extracted component: in particular, the inversion of the extracranial component of the REG II at Ce-Ce' suggests the existence of an electrode arrangement, probably in the neighboring of Cc-Cc' or Cd-Cd', from which a pure intracranial component could be directly recorded; the intracranial components generated by the model are morphologically similar to each other and to those recorded directly at Cc-Cc' and Cd-Cd'; their morphologies coincide with those reported in previous works in which the cerebral impedance changes were measured using intracranial electrodes (Laitinen, 1968); and, finally, the temporal derivative of the extracted component is similar to a typical transcranial Doppler trace.

The intracranial origin of the extracted component can be verified by means of the next experiment, which was carried out on volunteers. In this case, the current electrodes were placed at Cc and Cc' as shown in Figure 1, whereas the REG II pickup electrodes were attached to C5 and C6. In addition to recording the ECG signal for synchronization purposes, four photoplethysmographic sensors were fixed to the scalp surface at the frontal, occipital, and both temporal areas. The sensors consisted of a paired infrared transmitter and receiver, which provides an electronic signal that pulsates synchronized to the cardiac cycle in the same way as the scalp blood volume does.

Once the electrodes and sensors were placed, REG I, REG II, ECG and the four photoplethysmographic signals were acquired during two minutes. Then, a pneumatic cuff was inflated around the volunteer's head until the four photoplethysmographic signals turned flat. This means that the scalp blood flow has been arrested. Subsequently, all the channels were acquired for other two minutes. Like before, signal-to-noise ratio was improved by averaging the acquired data time-locked to the heartbeat as described above.

Preliminary results are summarized in Figure 4. This figure shows, overlapped, the intracranial REG component extracted from the first recording, and the REG I signal obtained when the cuff was inflated, both normalized to unit variance. As it was expected, REG I amplitude fell dramatically during the experi-

Figure 2. Set of five averaged REG waveforms recorded from the subject and normalized to unit variance; (a) REG I traces and time interval during which REG I impedance falls; (b) overlapped REG II traces and previous time interval transferred from A; positive Y axis represents increments in blood volume, and negative Y axis increments in impedance.

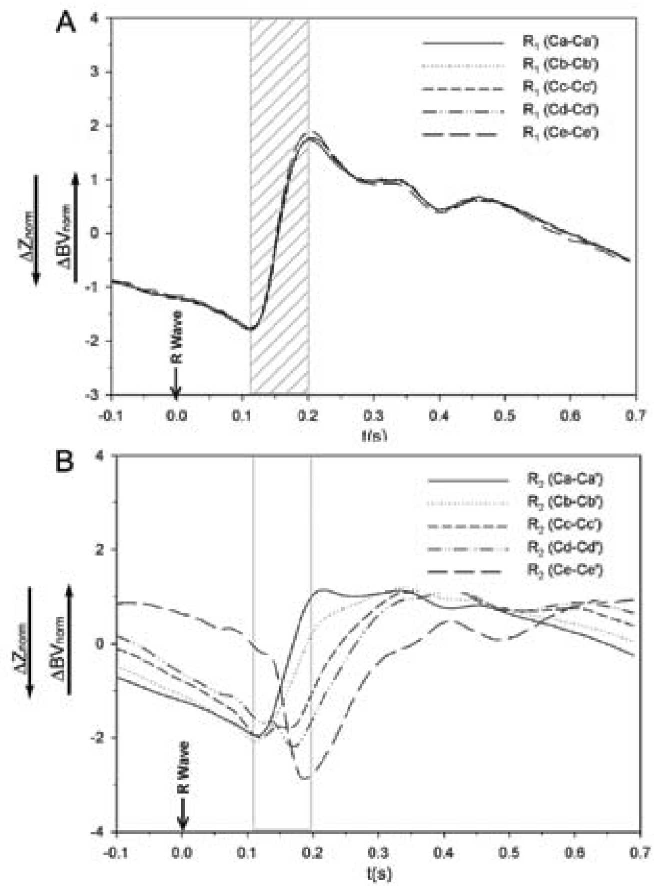


Figure 3. Intracranial components derived from each pair of REG signals by the proposed separation method. Boxed time interval has been transferred from that shown in figure 2A. Positive Y axis represents increments in blood volume and negative Y axis increments in impedance.

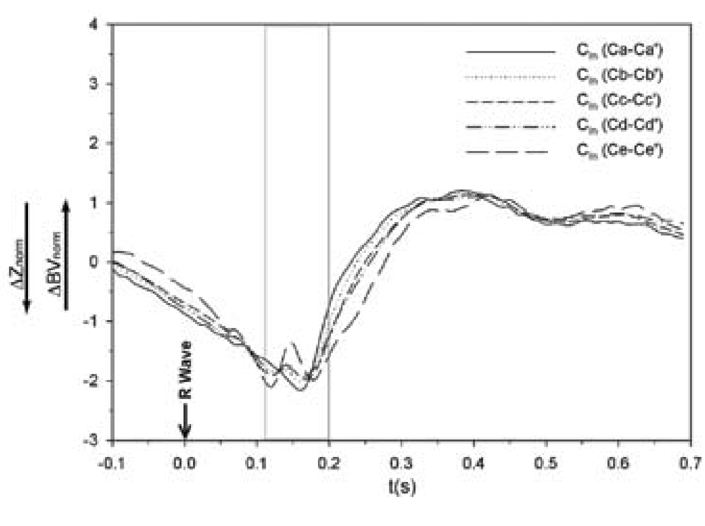
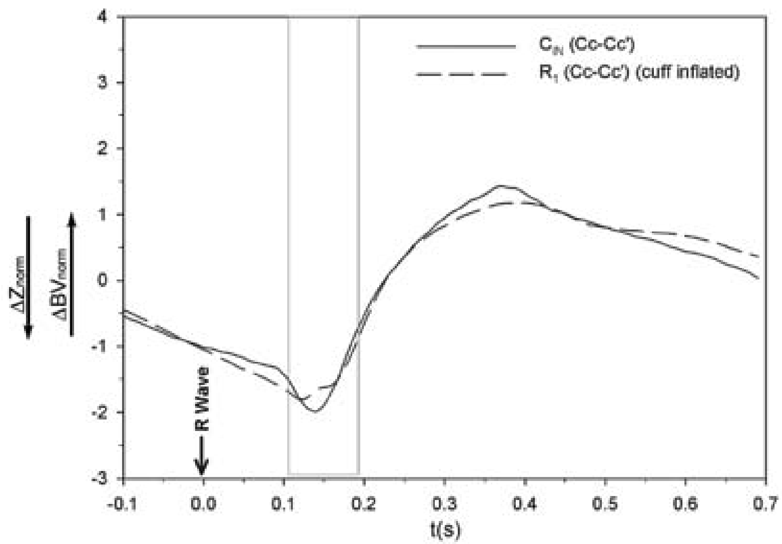


Figure 4. Intracranial component extracted by the separation method from a pair of REG signals ($C_{IN}(t)$) under normal conditions, and REG I recorded during scalp blood flow arrest forced by a pneumatic cuff ($R_1(t)$); boxed time interval has been transferred from that shown in Figure 2a; positive Y axis represents increments in blood volume, and negative Y axis increments in impedance.



ment when the scalp blood flow was arrested, since the cuff cancelled mechanically the extracranial REG component. Therefore, the residual REG I signal can be attributed to a nonextracranial origin. On the other hand, both waveforms shown in Figure 4 are practically identical to each other, and none of them has signs of the rising edge of the $R_1(t)$ trace impedance. This result suggests that the REG component extracted by the described method is of a true intracranial origin. Finally, since the measured variable is the intracranial impedance changes associated with the cardiac cycle, further analysis is required to find suitable REG signal parameters for application in the clinical practice.

FUTURE TRENDS

An effective separation of the two REG components requires to impose at least two constrains. The separation procedure described here is based on the physiological differences between the intra and extracranial blood flow, and assumes an exclusive extracranial origin of REG I. Although these conditions have proved to be effective, some other boundary conditions can be defined to improve the algorithm. In this sense, some other function cost could yield better results. Therefore, further research on these aspects would be needed.

Indeed, although the intracranial REG component could be completely separated from the raw signal, it is necessary to identify the appropriate parameters of the intracranial component that could be useful in the diagnosis and treatment of neurological diseases.

CONCLUSION

Research on REG was neglected, because the signal was strongly disrupted, and even buried, by an inherent extracranial artifact. Recent contributions to REG knowledge allow understanding its physical basis, and explaining some previous paradoxical results. The separation method of the intra and extracranial REG components described here, along with the preliminary results presented, brings some light to the classical problem of the REG, and makes it possible to set a new approach for further research on REG oriented to the clinical practice.

ACKNOWLEDGMENT

This work was supported by grant PI04/0303 from the Instituto de Salud Carlos III (Fondo de Investigación Sanitaria) in the framework of the 'Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica (I+D+I)' (SPAIN).

REFERENCES

Basano, L., Ottonello, P., Nobili, F., Vitali, P., Palavicini, F.B., Ricca, B., Prastaro, T., Robert, A., & Rodríguez, G. (2001). Pulsatile electrical impedance response from cerebrally dead adult patients is not a reliable tool for detecting cerebral perfusion changes. *Physiological Measurement*, *22*(2), 341–349.

Cardoso, J.F. (1998). Blind signal separation: Statistical principles. In *Proceedings of the IEEE* (Vol. 86, No. 10, pp. 2009–2025).

Geddes, L.A., & Baker, L.E. (1989). *Principles of applied biomedical instrumentation*. New York: Wiley.

Hadjiev, D. (1972). Impedance methods for investigation of cerebral circulation. *Progress in Brain Research*, *35*, 25–85.

Hyvärinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis*. New York: Wiley.

Laitinen, L.V. (1968). A comparative study on pulsatile intracerebral impedance and rheoencephalography. *Electroencephalography and Clinical Neurophysiology*, *25*(3), 197–202.

Lang, E.W., & Chesnut, R.M. (1994). Intracranial pressure. Monitoring and management. *Neurosurgery Clinics of North America*, *5*(4), 573–605.

Lifshitz, K. (1970). Electrical impedance cephalography (rheoencephalography). In M. Clynes, & J.H. Milsum (Eds.), *Biomedical Engineering Systems* (pp. 21–64). New York: McGraw-Hill.

Masucci, E.F., Seipel, J.H., & Kurtzke, J.F. (1970). Clinical evaluation of "quantitative" rheoencephalography. *Neurology*, *20*(7), 642–648.

Perez, J.J., Guijarro, E., & Barcia, J.A. (1999). Quantification of sensitivity to conductivity changes in bipolar rheoencephalography. In *Proceedings of the First Joint BMES/EMBS Conference* (Vol. 1, pp. 455).

Perez, J.J., Guijarro, E., & Barcia, J.A. (2000). Quantification of intracranial contribution to rheoencephalography by a numerical model of the head. *Clinical Neurophysiology*, *111*(7), 1306–1314.

Perez, J.J., Guijarro, E., & Barcia, J.A. (2004). Influence of the scalp thickness on the intracranial contribution to rheoencephalography. *Physics in Medicine and Biology*, *49*(18), 4383–4394.

Perez, J.J., Guijarro, E., & Sancho, J. (2005). Spatiotemporal pattern of the extracranial component of the rheoencephalographic signal. *Physiological Measurements*, *26*(6), 925–938.

Perez, J.J., Guijarro, E., Sancho, J., & Navarré, A. (2006). Extraction of the intracranial component from the rheoencephalographic signal: A new approach. In *Proceedings of the 28th IEEE – EMBS Annual International Conference* (Vol. 1, pp. 6064–6067).

Perez-Borja, C., & Meyer, J.S. (1964). A critical evaluation of rheoencephalography in control subjects and in proven cases of cerebrovascular disease. *Journal of Neurology, Neurosurgery, and Psychiatry*, *27*, 66–72.

Polzer, K., & Schuhfried, F. (1950). Rheographische untersuchungen am schädel. *Wiener Zeitschrift fur Nervenheilkunde und deren Grenzgebiete*, *3*, 295–298.

Seipel, J.H. (1967). The biophysical basis and clinical applications of rheoencephalography. *Neurology*, *17*(5), 443–451.

KEY TERMS

Bipolar Rheoencephalogram (REG I): Pulsatile impedance changes of the head related to the cardiac cycle that are measured by injecting a small current between two electrodes attached to the scalp surface and recording the electric potential difference generated between them.

Blind Source Separation: A part of Signal Processing that studies algorithms to extract a set of unknown signal sources from the analysis of a set of observations, each one of which is a mixture in unknown proportions of the unknown signal sources.

Electrical Impedance: Measure of the opposition to the flow of a sinusoidal alternating electric current.

Extracranial REG Component: Part of a REG signal that is caused by the pulsatility character of the blood flow in the extracranial tissues.

Intracranial REG Component: Part of a REG signal that is caused by the pulsatility character of the blood flow in the intracranial tissues.

Photoplethysmography: Technique to estimate the pulsatile changes of the blood volume in a tissue by means of the measurement of the light reflected on or transmitted through that tissue.

Rheoencephalography: Electromedical technique used to assess the cerebral blood flow (CBF) by noninvasive electrical impedance methods using electrodes attached to the scalp surface.

Tetrapolar Rheoencephalogram (REG II): Pulsatile impedance changes of the head related to the cardiac cycle that are measured by injecting a small current between two electrodes attached to the scalp surface and taking the electric potential difference between another pair of electrodes.

EONS an Online Synaptic Modeling Platform

Jean-Marie C. Bouteiller
University of Southern California, USA

Michel Baudry
University of Southern California, USA

Theodore W. Berger
University of Southern California, USA

INTRODUCTION

The nervous system is composed of billions of interconnected neurons. These neurons communicate with each other primarily through fast chemical synapses. Such category of synaptic transmission involves (i) the conversion from electrical to chemical signal at the presynaptic membrane; (ii) the chemical transmission (diffusion of neurotransmitter) across the synaptic cleft; and (iii) the conversion from chemical to electrical signal at the postsynaptic membrane. Chemical synapses, although representing the smallest unit of communication between two neurons in the nervous system, comprise a complex ensemble of mechanisms. Understanding these mechanisms and the way synaptic transmission occurs and is regulated by activity is critical for our comprehension of central nervous system (CNS) functions in general and learning and memory in particular.

The objective of the work presented here is to develop a computational tool designed specifically for the study of the complex mechanisms that occur at chemical synapses. This tool should enable students and researchers (biologists with very little training in computational biology, as well as experienced modelers) worldwide to study the roles of diverse parameters that impact synaptic transmission from presynaptic to postsynaptic depolarization in an integrated modeling platform, using an easy, user-friendly and modular graphical interface.

BACKGROUND

Computational neuroscience has since its inception been mainly focusing on capturing the essential features of biological systems at multiple spatiotemporal scales, from membrane currents, protein and chemical coupling to network oscillations and learning and memory. These computational models are used to test hypotheses that can be directly verified by current or future biological experiments.

Recent technological advances have provided a better understanding of the mechanisms involved in synaptic transmission and integration; in particular, calcium imaging (Oertner, Sabatine, Nimchinsky, & Svoboda, 2002) and more accurate electrophysiological measurements have provided precise information on calcium concentration and dynamics (Oakley, Schwindt, & Crill, 2001), density, and types of channels and receptors, as well as different molecular pathways.

Likewise, the collection of tools available for computational approaches in biology has also grown significantly in recent years. This collection contains (as a nonexhaustive list): Neuron, ECELL, BioSpice, MCell, Genesis, and Virtual Cell (more information is available on their respective Web sites). Among those tools, modeling platforms have been developed to study complex cellular mechanisms. Some approaches focus on studying very specific mechanisms in simple geometries (Bennett, Farnell, & Gibson, 2000; Matveev, Zucker, & Sherman, 2004), while others study cellular mechanisms in complex and up to three-dimensional geometry (MCell) and offer a much more general framework (virtual cell).

Chemical synapses have been the subject of extensive research very early in the history of neuroscience (Rall, 1974). They have been shown to play a significant role in learning and memory; thus, synaptic structure is a critical determinant of synaptic properties, while temporal changes in postsynaptic calcium concentration and CaMKII or phosphatase activation generated by various stimulation conditions are believed to be good predictors for the induction of Long Term Potentiation

(LTP) or Long-Term Depression (LTD). Furthermore, chemical synapses also constitute a primary target for a multitude of drugs used in pathologies affecting the central nervous system (such as AMPAkines or NMDA receptor antagonist in the case of Alzheimer's disease).

The parameters that affect the efficacy of chemical synapses are numerous. A nonexhaustive list would include presynaptic mechanisms (calcium channels kinetics and distribution, presynaptic calcium binding and buffering mechanisms, kinetics of vesicle cycling, calcium extrusion from the terminal, and so forth), synaptic mechanisms (diffusion of neurotransmitters in the synaptic cleft, reuptake mechanisms, width of the cleft, relative position of the release site with respect to the receptors, and so forth), and postsynaptic mechanisms (receptor affinity and distribution, channel-gated kinetics, second-messenger pathways, postsynaptic spine geometry, and so forth).

Given the complexity of the mechanisms involved and the importance of chemical synaptic transmission in the central nervous system, we believe it is of paramount importance to develop a computational platform to specifically study the contribution of different parameters to synaptic transmission and regulation.

MAIN FOCUS OF THE ARTICLE

Requirements, Architecture of EONS

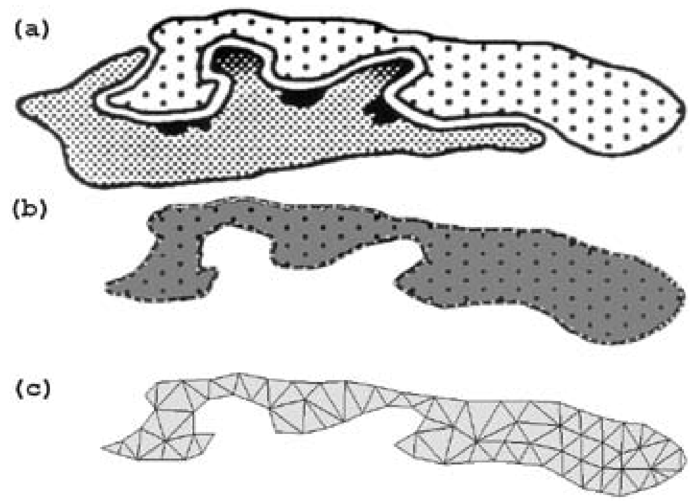
The EONS synaptic modeling platform is an integrated modeling framework that allows the user to specify in a structured environment using a graphical user interface the characteristics of the synaptic elements one wants to study. From a computational standpoint, EONS consists of a Java WebStart application that can be downloaded from the World Wide Web (http://eons.no-ip.info). The application communicates with a central database in which various models and elements are stored. Hence, users can save and retrieve models and/or parts of these models.

The platform contains models, structures, elements, reactions, and simulations. Models are entities in which all simulated components are defined. They represent the whole system one is interested in. This system can be an entire synapse, or simply a calcium channel. Structures are containers for modeling elements. A container can be conceptual (one can think of it as a dimensionless toolbox in which several elements can be grouped) or can become a physical container when associated to a specific geometry. It then adopts the dimensions of the two-dimensional mesh with which it is associated. A model can contain several structures (e.g., presynaptic, cleft, postsynaptic, and so forth). Elements are unitary entities that can be added or removed from the rest of the model. An element contains parameters which can be constant or variable. As an example, an element can be a calcium channel, a calcium pump, or a postsynaptic receptor. Elements interact with their environment. These interactions are described as Reactions in EONS. Reactions represent a set of mathematical expressions that describe the events or interactions occurring in the system. Simulations represent the actual in-silico experiment. A simulation can solely be run on an entire model. Results of a given simulation are observed in the form of graphs and an array of values for every time step of the simulation.

Mathematics

The description of biological models and their subsequent computation often requires mathematical tools. EONS uses the linear algebra module available in jScience (Dautelle, 2006) to generate matrices and calculate the values of its coefficients. This is used in particular for the diffusion process using Finite Element Method (FEM). In the same way, modeling requires the users to define the behavior of the elements they intend on modeling. To do so, EONS allow users to enter their own sets of mathematical expressions and parses these expressions using a mathematical expression parser called Java Math Expression Parser (JEP) (Funk & Morris, 2006). JEP is a Java package for parsing and evaluating mathematical expressions. It supports user-defined variables, constants, and functions. A number of common mathematical functions and constants are included (see reference Web site). To complement the functions available in JEP, a set of mathematical methods has also been implemented to serve the specific needs of physiological and cellular modeling which require the use of different solving methods for ordinary differential equations. Hence, numerical methods such as forward Euler, backward Euler, and Runge-Kutta 2nd and 4th order have also been implemented.

Figure 1. Creation of the 2D mesh. (a) Initial raster contour drawn from the EM picture. (b) Creation of the to-scale contours (white dots around the presynaptic terminal. (c) Output of mesh2D mesh generation program that is used as an input to EONS. The triangles form a mesh through out the geometry, where each vertex in called a node.



Modeling Synaptic Elements

EONS already contains a library of synaptic models. As of the redaction of this manuscript, those elements include calcium channels (L, N and T type channels) as well as postsynaptic elements such as AMPA and NMDA receptor channels. This library can easily be expanded to contain other elements such as calcium pumps, deterministic vesicular release, exchange and reuptake mechanisms, and so forth. The AMPA receptor channels modeled are the 5-states receptor introduced in (Patneau & Mayer, 1991) and the 7-states described in Jahn, Butler, and Franke (1998). The NMDA receptor channel modeled is the 11 states model presented in (Lester, Tong, & Jahr, 1993).

As an example, Figure 2 illustrates the results obtained during the simulation of the 5-states AMPA receptor channel. The input to the model is a pulse of glutamate. In this case, no diffusion is applied to the model (results of the diffusion process will be explained in details in the next section). The model is stated as a set of 5 ordinary differential equations, each represented in EONS using the syntax presented in section (b) of Figure 2.

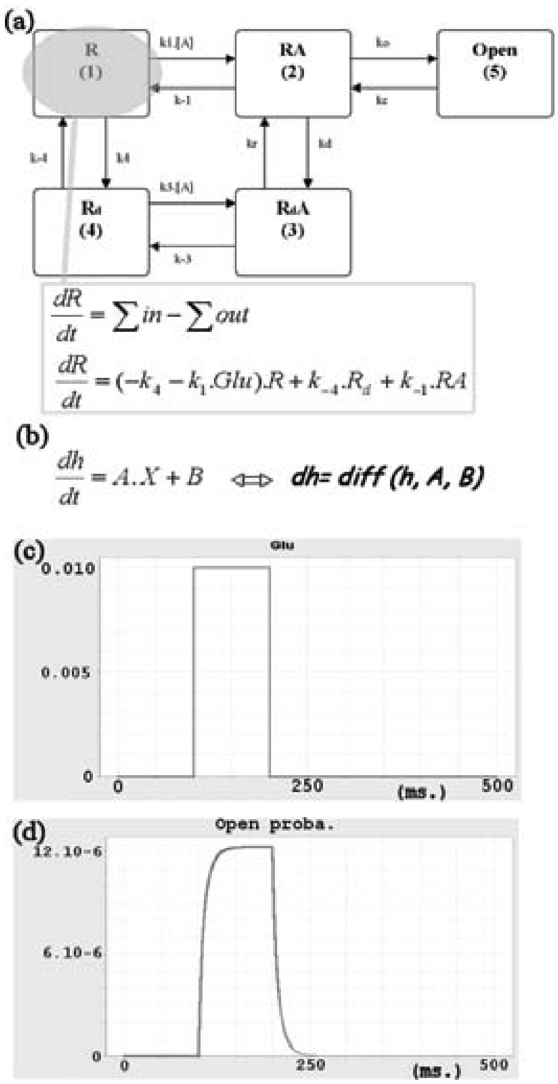
Molecular Diffusion: From Image to Mesh

In most current approaches to modeling relevant to our problem, the regions of interest are often simplified with basic geometrical shapes (Smith, 1996). Hence, information that depends on the spatial organization of molecules and cellular organelles cannot be accommodated.

In EONS, we propose to retain this critical information and to model diffusion in realistic complex geometry using finite element method. The first step in this process consists in tracing the contours of the structure of interest using real geometrical information (i.e., electron microscopy picture) and scaling them to the original size (generally in nanometers). Then, we use an external C program (Karamete, 1992) to generate the mesh which will be uploaded into EONS. These steps are illustrated using a perforated presynaptic terminal (Geinisman, 1993) in Figure 1.

The utilization of this finite element method allows us to discretize the computational domain by approximating the continuum problems (Ghaffari-Farazi, Liaw, & Berger, 1999). The diffusion equation in the continuum structure is:

Figure 2. 5-states AMPA receptor channel modeling in EONS. (a) Kinetic schema of the receptor. The ordinary differential equation generated for state R is represented. (b) Ordinary differential equation and its corresponding syntax in EONS. (c) Input of the model: Pulse of glutamate (10 µM. for a duration of 100 msec during a 500 msec simulation with a 0.01 msec time step). (d) Output of the model: open probability of the associated channel.



$$\frac{\partial c(x,y,t)}{\partial t} = D_c.\{\frac{\partial^2 c(x,y,t)}{\partial x^2} + \frac{\partial^2 c(x,y,t)}{\partial y^2}\} - B(t)$$

where c(x,y,t) represents the concentration function, Dc is the diffusion coefficient (for free calcium ions, its value is 2.2 . 10-6 cm2/sec in the cytoplasm (Jackson & Redman, 2003)). B(t) is the buffering process when it is assumed to be immobile and uniformly distributed.

In the presynaptic terminal, the process of calcium binding to buffers occurs in a much faster time scale than diffusion, and local equilibrium is rapidly reached between free and bound calcium ions (Crank, 1975). Therefore, the above equation can be approximated by:

$$\frac{\partial c(x,y,t)}{\partial t} = \frac{D_c}{1+\beta}.\{\frac{\partial^2 c(x,y,t)}{\partial x^2} + \frac{\partial^2 c(x,y,t)}{\partial y^2}\}$$

where β is the ratio of bound to free calcium in the cytoplasm.

To discretize the computational domain, we define a shape function N such that the value of the concentration is equal to the discrete values of [c] on the points of the mesh, which in the case of one triangle yields:

$$c = \sum_{i=1} C_i.N_i(x,y) = \begin{bmatrix} C_1 & C_2 & C_3 \end{bmatrix}.\begin{bmatrix} N_1 & N_2 & N_3 \end{bmatrix}^T = C.N^T$$

The utilization of Galerkin principle applied to the diffusion equation yields:

$$Dc.\int_S \nabla c.N.n.dS$$
$$- D_c.[c]\int_V (\frac{\partial N^T}{\partial x}\frac{\partial N}{\partial x} + \frac{\partial N^T}{\partial y}\frac{\partial N}{\partial y})dx.dy$$
$$-[\dot{c}].\int_V N^T.N.dx.dy \quad = 0$$

Solving this system gives us the values of $[\dot{c}]$, values of the gradient of concentration in the mesh and hence the values of the concentration at every node at the next iteration.

Modeling Diffusion in EONS

Simulating diffusion in complex arbitrary geometrical shapes can easily be done in EONS. Accessing the diffusion method is done using the following statement:

$$\text{var}\,iable = FEM\,(\text{"}mesh\_name\text{"},\ initial\ concentration,\ diffusion\ coefficient)$$

The effect of this mathematical expression is twofold: (i) it initializes a matrix (named "Variable") according to the specified mesh in which every cell contains the value of the initial concentration provided as second parameter, and (ii) it calculates the diffusion throughout the simulation of a molecule for which the diffusion coefficient in the medium is given as third parameter. In order for a specific element to have an effect on the concentration at a specific location, one must apply the function:

$$inject(mesh\_matrix,\\ \quad position\ in\ the\ mesh,\\ \quad \mathrm{var}iable\ to\ be\ added)$$

For a calcium channel, for example, the increase in concentration observed when the channel is in the open state is added to the mesh matrix at the position where the channel is located. To determine the effect of geometry on synaptic function, it is important to have access to the value of the concentration of a molecule at every position throughout the mesh. To do so, it is possible to simulate a recording electrode at any position on the mesh using the function:

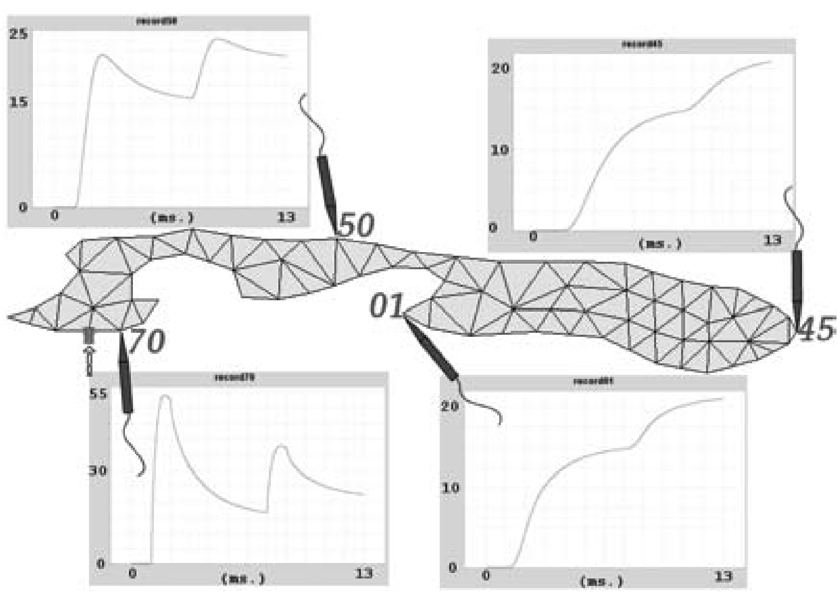
$$record(mesh\_matrix,\\ \quad position\ in\ the\ mesh)$$

Initial results show that the nonoptimized computation time needed to calculate the diffusion in a presynaptic terminal such as the one presented in Figure 3 is equal to 10 minutes for a 12 msec simulation on a 3.2 GHz Pentium workstation (with a time step equal to 0.1 μsec).

FUTURE TRENDS

The current version of the synaptic modeling platform presented here can handle a large number of models as well as reaction-diffusion processes. However, many further enhancements are planned. There is currently no way to specify a nonuniform population of molecules in a physical space, such as calcium-binding molecules in the presynaptic terminal (Matveev et al., 2004). The current library of models will be extended to include more complex molecular pathways involving in particular G protein coupled receptors, second messenger pathways, and endogenous feedback from postsynaptic elements to presynaptic terminals. The

Figure 3. Virtual simulated electrodes and their corresponding plots. These "electrodes" record the concentration of calcium at different locations throughout the presynaptic terminal (mesh containing 86 triangles) following a paired-pulse depolarization (0 mV and 1 msec long with a 6 msec interval) with calcium entering the terminal through one N-type calcium channel with no buffering mechanisms.



addition of more pathways will provide opportunities to study the synaptic system as a whole, but also allow us to simulate pathologies and the effect of drugs on synaptic transmission. In the same way, work will be done with respect to compatibility between EONS and other modeling platforms. This will enable the interchange of models between modeling tools, an example of which would consist in building a realistic neuron with the Neuron platform while incorporating synapses modeled with precise dynamics using EONS. Strong modularity between computational tools certainly bodes well for the future of the field of computational biology.

In its current state, EONS provides both researchers and students with a well-defined environment for testing specific hypotheses about the dynamics of multiple presynaptic mechanisms, such as the type of Ca2+ channels, buffering mechanisms, Ca2+ pumps, diffusion coefficients, binding proteins, and so forth, as well as several key postsynaptic mechanisms, such as the kinetics of AMPA and NMDA receptor-channel subtypes. In addition, one of the unique features of EONS with respect to other neural modeling tools is its ability to preserve the effect of synaptic geometry, for example, the relative locations of presynaptic Ca2+ channels and neurotransmitter release sites, the relative locations of presynaptic release sites and postsynaptic receptor-channels, and the relative numbers and locations of postsynaptic receptor-channels, while integrating mechanisms that are both pre- and postsynaptic.

CONCLUSION

Biological organisms are of extreme complexity. Modeling such complexity is not trivial: first, it involves the precise knowledge of all or at least a large number of the mechanisms involved. Then, once the details pertaining to those mechanisms are understood, one must model them in the most accurate manner. The amount of work involved in these tasks is tremendous and years to come will see many significant advances in the field of neurocomputation. The first challenge will come from the large number of elements, the even higher number of interactions between them and the complexity of interrelated pathways. The second challenge will come from the multihierarchical structure of organisms which span from a multitude of elements ranging from the nanoscopic level to an elaborate structured macroscopic level. Finding solutions to integrate these two levels of complexity will constitute the key to accurately model complex systems.

Although consisting of a single unit of communication between two neurons, chemical synapses are the siege of a very large number of phenomena and molecular interactions which without any doubts help shape the way we perceive, learn and remember. The research presented here provides a representative approach to the problem of modeling chemical synapses. We focused our attention primarily on the EONS synaptic modeling platform as a novel integrated framework containing both presynaptic terminal and postsynaptic density with modular geometry. We described the elements currently modeled and the main characteristics of the approach and presented future research directions that will help extend the features of the platform and further strengthen our knowledge of chemical synaptic transmission and of the mechanisms involved in learning and memory. We also expressed our belief for the need of integrated modeling environments to allow the field of neurocomputation to study beyond unitary mechanisms and take into account interrelated phenomena that code much of the complexity encountered in the central nervous system.

REFERENCES

Bennett, M. R., Farnell, L., & Gibson, W. G. (2000). The probability of quantal secretion near a single calcium channel of an active zone. *Biophysical Journal, 78*, 2201-2221.

Crank, J. (1975). *The mathematics of diffusion* (2nd ed.). Oxford: Clarendon Press.

Dautelle, J.-M. (2006). *JScience*. Retrieved February 14, 2008, from http://jscience.org/

Funk, N., & Morris, R. (2006). *JEP: Singular systems.* Retrieved from http://www.singularsys.com/jep

Geinisman, Y. (1993). Perforated axospinous synapses with multiple, completely partitioned transmission zones: Probable structural intermediates in synaptic plasticity. *Hippocampus, 3*, 417-434.

Ghaffari-Farazi, T., Liaw, J., & Berger, T. (1999). Consequence of morphological alterations on synaptic function. *Neurocomputing, 26*, 17-27.

Jackson, M. B., & Redman, S. J. (2003). Calcium dynamics, buffering, and buffer saturation in the boutons of dentate granule-cell axons in the hilus. *The Journal of Neuroscience, 23*(5), 1612.

Jahn, K., Butler, J., & Franke, C. (1998). Kinetics of AMPA-type glutamate receptor channels in rat caudate-putamen neurones show a wide range of desensitization but distinct recovery characteristics. *European Journal of Neuroscience, 10*, 664-672.

Karamete, B. K. (1992). *2D constrained Delaunay unstructured mesh generator: Mesh2d.*

Lester, R. A. J., Tong, G., & Jahr, C. E. (1993). Interactions between the glycine and glutamate binding sites of NMDA receptor. *Journal of Neuroscience, 13*, 1088-1096.

Matveev, V., Zucker, R. S., & Sherman, A. (2004). Facilitation through buffer saturation: Constraints on endogenous buffering properties. *Biophysical Journal, 86*, 2691-2709.

Oakley, J., Schwindt, P., & Crill, W. (2001). Initiation and propagation of regenerative action potentials in dendrites of layer 5 pyramidal neurons. *J. Neurophysiology, 86*, 503-513.

Oertner, T. G., Sabatine, B. L., Nimchinsky, E. A., & Svoboda, K. (2002). Facilitation at single synapses probed with optical quantal analysis. *Nature Neuroscience, 5*, 657-664.

Patneau, D. K., & Mayer, M. L. (1991). Kinetic analysis of interactions between kainate and AMPA: Evidence for activation of a single receptor in mouse hippocampal neurons. *Neuron, 6*, 785-798.

Rall, W. (1974). Dendritic spines, synaptic potency and neuronal plasticity. In C. D. Woody, K. A. Brown, T. J. Crow & J. D. Knispel (Eds.), *Cellular mechanisms subversing changes in neuronal plasticity* (pp. 13-21). Los Angeles: Brain Information Services.

Smith, G. D. (1996). Analytical steady-state solution to the rapid buffering approximation near an open Ca2+ channel. *Biophysical Journal, 71*, 3064-3072.

KEY TERMS

AMPA, NMDA, mGluR Receptors: Different subtypes of glutamate receptors. AMPA and NMDA receptors are ionotropic (directly linked to an ion-channel) while the mGluR family of receptors is metabotropic (indirectly linked with ion-channels through signal transduction mechanisms).

Chemical Synapses: Chemical synapses are specialized junctions through which cells constituting the nervous system signal to one another and to non-neuronal cells such as muscles or glands. They represent key elements regulating communication between neurons in neural circuits.

Finite Element Method: Finite element method allows the calculation of diffusion in complex spaces by discretizing the continuous computational domain. The diffusion equation is then solved as a series of differential equations in each discretized element of the domain.

Glutamate: Glutamate is the most abundant excitatory neurotransmitter in the mammalian central nervous system. It is stored in vesicles and its release is triggered by nerve impulses.

Java Webstart Application: Java is an object-oriented programming language developed by Sun Microsystems. It allows the development of Webstart applications that can be downloaded from the Internet and run in a platform-independent manner.

Long Term Potentiation: Long Term Potentiation is generally considered to represent a cellular mechanism underlying certain forms of learning and memory. Long term potentiation consists of an increase in the strength of a chemical synapse that lasts from minutes to several days or weeks following a series of short, high-frequency electrical stimuli.

Postsynaptic Density: The postsynaptic density is a membrane specialization that concentrates neurotransmitter receptors and associated proteins to respond rapidly to the presence of neurotransmitter in the space separating the two neurons (referred to as the synaptic cleft).

Ethical Decision-Making in Biomedical Engineering Research

Monique Frize
Carleton University, Canada
University of Ottawa, Canada

Irena Zamboni
Carleton University, Canada

INTRODUCTION

To be ethical and professional are terms that are synonymous with being an engineer. The work of engineers frequently affects public safety and health, and can influence business, and even politics. Professional Engineering Associations provide ethical guidelines so that engineers will know how to avoid misconduct, negligence, incompetence, and corruption, which could lead to formal complaints and discipline. Knowledge about ethical decision-making guides engineers facing complex and difficult moral dilemmas (Andrews, 2005, pp. 46). Biomedical engineers doing research and development will undoubtedly be involved in projects that impact humans and/or animals, and thus must be informed on all aspects of ethics that guide such research. They should be particularly aware of the specific guidelines of the institution where the work is to be carried-out and be familiar with the application process to obtain a certificate, allowing the research to proceed. There is clearly a need to guide biomedical engineering students and practitioners in performing a balanced analysis of difficult questions and issues, while respecting societal values that may differ greatly from their own (Frize, 1996; Frize, 2005; Saha & Saha, 1997; Wueste, 1997). There exists a number of articles discussing biomedical engineering and ethics specifically aimed at clinical engineers (Goodman, 1989; Saha & Saha, 1986). These are helpful readings for anyone involved in biomedical research or clinical engineering.

RELEVANCE OF ETHICAL THEORIES FOR DECISION-MAKING IN BIOMEDICAL ENGINEERING RESEARCH

Morality is defined as what people believe to be right and good, and the reasons for it. There are typical rules of conduct describing what people ought and ought not to do in various situations. Ethics is the philosophical study of morality. It is a rational examination into people's moral beliefs and behavior, the study of right and wrong, of good and evil in human conduct.

Several ethical theories exist, but some are more relevant for decision-making for engineers than others. For example, theories of Subjective Relativism and Cultural Relativism have limited utility in ethical decision-making, as they encourage decisions based on individual or cultural perspectives. The Divine Command Theory, based on particular religious beliefs which can vary from one religion to the other, also has limited value for this type of decision-making. Theories such as Kantianism, Act Utilitarianism, Rule Utilitarianism, Social Contract, Rights Theory, and Rawl's Theory of Justice appear to be more helpful for decision-making related to biomedical research.

Subjective Relativism (SR): SR is a theory where different individuals or groups of people can have completely opposite views of a moral problem, and both can be considered right. Persons decide right and wrong for themselves. The case for supporting this point of view is that well-meaning and intelligent people can have opposite opinions about moral issues, but the line between doing what you think is right and doing what you want to do is not sharply drawn. For example, we can rationalize bad behavior or allow persons to decide right and wrong for themselves. SR makes no moral distinction between the actions of different people, as, for example, between Hitler and Mother Teresa. SR is based on the idea that there are no universal norms. However, as we see in several of the theories described further, there are some universal moral principles and norms. Examples are: not to kill, not to break promises, to be fair and honest. The self-defeating nature of SR does not make it a workable theory for ethical decision-making in biomedical engineering research (Quinn, 2005, pp. 53–55).

Cultural Relativism (CR): CR is the ethical theory where the meaning of right and wrong rests with a society's current moral guidelines. Folkways get institutionalized into these guidelines. This is based on the fact that in the distant past, survival was a group activity. This can still be true in some instances, as in situations of security against crime, war, or terrorism. CR is folkways turned into customs, traditions which are uniform, universal for the group. But with time, they can become more and more arbitrary and imperative. Morality of the group is the sum of the taboos and prescriptions in the folkways by which right conduct is defined for a particular time. For example, 10,000 years ago, most activity was food gathering. Today, population explosion has created huge environmental problems which can destroy the planet. So we need different rules of conduct in various eras. Anthropologists have documented important differences on what is proper conduct in various epochs and regions. However, just because two societies have different views about right and wrong does not imply that they ought to have different views. Perhaps one society has "good" guidelines, and another has "poor" ones. CR does not explain how an individual determines the moral guidelines of a particular society. Another consideration is the concept of human rights. If CR allows acts that fundamentally breach human rights, then this should not be tolerated. This leads to the next objection to this theory: CR does not explain how moral guidelines evolve. CR suggests there are no universal moral guidelines, and gives tradition more weight in ethical evaluations than facts or reason; it does not provide a framework for reconciliation between cultures in conflict. The value of each society can lead to actions that harm the other, yet cultural relativism says each society's moral guidelines are right. CR does not provide a way for the two sides to find common ground. All societies, in order to maintain their existence, must have a set of core values; for example, caring for helpless newborn babies; telling the truth; prohibition of assault, rape, and murder. CR has significant weaknesses as a tool for constructing ethical evaluations for a diverse audience (Quinn, 2005, pp. 55–59).

Divine Command Theory (DCT): Judaism, Christianity, and Islam are religions based on a god, and DCT is based on the idea that good actions are those aligned with the will of God; bad actions are those contrary to the will of God; we owe obedience to our Creator, and God is all-good and all-knowing, the ultimate authority. There exist many holy books, and some of their teachings disagree with each other. It is unrealistic to assume that a multicultural secular society (separation of state and church) can adopt a religion-based morality. A society's moral guidelines should emerge from a secular authority. Moreover, some moral problems are not addressed directly in scriptures, as, for example, those arising from Internet or information technology. Even if an analogy was to be used, who interprets it? This becomes a subjective process. The fact that the ethical guidelines are not the result of a logical progression from a set of underlying principles is a significant obstacle. So the DCT is not a powerful tool for ethical debate in a secular society, and not a workable theory for our purpose of ethical decision-making in this context (Quinn, 2005, pp. 59–62).

Kantianism (Kant, 1724–1804): This theory, also referred to as the categorical imperative, pertains to actions that are universally considered to be good, and involve good will and duty. We are compelled to act in a certain way because of some moral rule. In his first formulation, Kant defines the categorical imperative: "*Act only from moral rules that you can, at the same time, will to be universal moral laws"* (Quinn, 2005, pp. 63). An example is: Do not make a promise with the intention of breaking it. If everyone broke promises, they would become meaningless. Kant's second formulation states: "*Act so that you always treat both yourself and other people as ends in themselves, and never only as a means to an end"* (Quinn, 2005, pp. 64). This aspect applies well to the question of self-interest in overly ambitious researchers, or to the fact that some researchers do not tell the whole truth to human subjects about the experiments they will perform on them. It is wrong for a person to use another; all interactions must be respectful of the humanity of others. Kantianism is rational; that is, we can use logic to explain the "why" behind a solution to an ethical problem. This theory produces universal moral guidelines that apply to all people for all times. All persons must be treated as moral equals; that is, people in similar situations should be treated equally. Kantianism provides a framework to combat discrimination. A weakness of this theory is that sometimes an action cannot be fully characterized by a single rule. For example, is it acceptable to steal food to feed starving children? This approach does not help to resolve a conflict between rules. How do we rank the importance of these rules? Kantianism does not allow any exceptions to moral laws. This theory mainly supports moral decision-making based on logical reasoning from facts and commonly held views. It is culture-neutral, and treats all humans as equals. It is a workable theory for ethical decision-making in cases where universal principles are concerned (Andrews, 2005, pp. 127–128,130; Quinn, 2005, pp. 62–67).

Act Utilitarianism (Act U): First formulated by Jeremy Bentham (1748–1832), and further developed by John Stuart Mill (1806–1873), Act Utilitarianism is based on the principle of utility. This theory is in sharp contrast to Kantianism: it states that an action is good if it benefits someone, and bad if it harms someone. This is also called the "greatest happiness principle." An action is right (or wrong) to the extent that it increases (or decreases) the total happiness of the affected parties. Happiness is defined as advantage, benefit, good, or pleasure; and unhappiness as disadvantage, cost, evil, or pain. We must weigh positive and negative impacts of an act, or the morality of an action. The focus is on the consequence of the actions. Anyone who can experience pleasure or pain is included in the calculations. The intensity, duration, certainty, propinquity, purity, and the extent of the impact must be compiled. This theory is pragmatic, quantitative to a certain extent, and comprehensive; it considers all elements of a situation. A shortcoming is that, when performing the utilitarian calculations, it is not clear where we draw the line (who is included, for what time period, and so on); yet, where the line is drawn can change the outcome of the evaluation, and can be a difficult task. Moreover, it is not practical to put so much energy into every single moral decision. One practical approach is to develop rules of thumb; for example, "it is wrong to lie." If a rule of thumb does not apply, then we can do all the calculations in detail to assess a unique situation or act. Contrary to Kantianism, Act U ignores our innate sense of duty; all that matters are the consequences of an action; the approach does not take into consideration the fact that people often act out of a sense of duty or obligation, as Kant suggests. Another weakness is that Act U cannot predict unintended consequences, so it applies mainly where consequences are known and controlled. Act U is an objective, rational, ethical theory that allows a person to explain why a particular act is right or wrong, and is a workable ethical theory to evaluate moral problems for biomedical engineers (Quinn, 2005, pp. 67–72).

Rule Utilitarianism (Rule U): John Stuart Mill's development of Rule U simplifies the multiple calculations needed for Act U. This approach holds that we ought to adopt moral rules which, if followed by everyone, will lead to the greatest increase in total happiness. It applies the Principle of Utility to moral rules, while Act Utilitarian applies the Principle of Utility to individual moral actions. Both Rule U and Kantianism are focused on rules, and some of the rules overlap. Both theories support the notion that rules should be followed without exception. However, the two theories derive moral rules in different ways. Rule U states that you follow a moral rule, because its universal adoption would result in greatest general happiness (consequences); whereas Kantianism states that you follow a moral rule, because it is in accord with the Categorical Imperative (motivation and duty). Performing the Rule U calculations is much simpler than for Act U. One weakness is that utilitarianism forces us to use a single scale (units) to evaluate completely different kinds of consequences. Moreover, this approach ignores the problem of an unjust distribution of good consequences. In summary, Rule U treats all persons as equals, and provides its adherents with the ability to give the reasons why a particular action is right or wrong. It is a workable theory for evaluating moral problems facing engineers. We should act so that the greatest amount of good is produced, and distribute the good as widely as possible (Andrews, 2005, pp. 126–127; Quinn, 2005, pp. 72–76).

Social Contract Theory (SC): Thomas Hobbes (1603–1679), in *Leviathan,* wrote that morality consists in the set of rules governing how people are to treat one another, that rational people will agree to accept for their mutual benefit, on the condition that others follow those rules as well. Similar to Kantianism, there are universal rules that can be derived through a rational process. But Kantianism is based on the universality of a rule, whereas SC is based on the benefits to the community, and recognizes the harm that a concentration of wealth and power can cause. It is framed in the language of individual rights, and explains why rational people act out of self-interest in the absence of common agreement. It also provides a clear ethical analysis of some important moral issues regarding the relationship between people and government. A weakness of the theory is that no one signed the social contract. Moreover, some actions can be characterized in multiple ways. SC theory does not explain how to solve a moral problem when the analysis reveals conflicting rights; it may even turn out to be unjust for people who are incapable of upholding their side of the contract, or for people who do not understand the rules (Quinn, 2005, pp. 76–84) .

Rights Theory (John Locke, 1632–1704): Locke's theory states that everyone has rights arising simply from being born; the right to life, maximum individual liberty, and human dignity are all fundamental rights, and other rights arise as a consequence of these. The difference with Kantianism is that for Locke, duty is a consequence arising from personal rights. The principles of Locke's theory are embedded in the "Charter of Human Rights and Freedoms" in many countries; this guarantees to all citizens fundamental freedom of conscience, religion, thought, belief, opinion, expression, peaceful assembly, and association. The legal right to life, liberty, and security of the person, and the right not to be deprived of these rights—except in accordance with principles of fundamental justice—are written in this law. This includes equality rights before and under the law, and the right to equal benefit and protection of the law. This theory is usable in ethical decision-making by engineers, especially with regards to protection of the public and of research subjects (Andrews, 2005, pp. 128–130).

Rawl's Theory of Justice (John Rawl, 1921–): In his *A Theory of Justice*, Rawl wrote that each person may claim a "fully adequate" number of basic rights and liberties, such as freedom of thought and speech, freedom of association, the right to be safe from harm, and the right to own property, so long as these claims are consistent with everyone else having a claim to the same rights and liberties. This resembles the Hobbes SC Theory and Locke's Rights Theory, except that it is stated from the point of view of rights and liberties, rather than from the view of moral rules. Any social and economic inequalities must satisfy two conditions: First, they are associated with positions in society that everyone has a fair and equal opportunity to assume; and second, they are to be to the greatest benefit of the least advantaged members of society (the difference principle). In reality, regarding the first principle, we cannot imagine everyone having the same decisional power and same wealth; it is a certainty that some people will have more power and/or money than others. The second principle is more likely to be applied; an example would be that the poor should pay less tax than the rich. Like Locke, Rawl reaffirms the right to life, liberty, and security of the person, and the right not to be deprived of these rights, except in accordance with principles of fundamental justice. This is a usable theory for decision-making by engineers (Quinn, 2005, pp. 78–80).

It is important to note that ethical decision-making is a dynamic process, as the issues and definitions change with country, historical era concerned, and the appearance of new ethical issues created by new research such as stem cells, designer babies, and cloning. Everyone must keep in mind an unfailing principle that applies to everyone, and for all times: Every human being, of all races, ethnic backgrounds, and sex are equal and must be given respect, dignity, and freedom from harassment and discrimination.

Other tools that can help engineers in their ethical decision making process are codes of ethics and guidelines. Principal codes are: the Hippocratic Oath, the Nuremberg Code, and the Declaration of Helsinki. The Hippocratic Oath is taken by doctors when they graduate, and sets out moral and ethical obligations to patients regarding the sanctity of human life, relief of suffering, and so on. (Illingworth & Parmet, 2006, pp. 173; Miles, 2004, pp. xiii–xiv). The Nuremberg Code, created as a result of the Nuremberg trials at the end of World War II, deals with principles of human experimentation and voluntary informed consent (Illingworth & Parmet, 2006, pp. 505). Understanding of both the Hippocratic Oath and the Nuremberg Code is essential for engineering disciplines involving work with human subjects. The World Medical Association (2004) published the Declaration of Helsinki as a statement of ethical principles to provide guidance to physicians and others when performing research involving human subjects, identifiable human material, or data (Illingworth & Parmet, 2006, pp. 506–509). Even the use of archived data without identifiers needs some form of clearance by ethics review committees. In addition, there are also the Code of Ethics for engineers in various States and Provinces, and the IEEE Code of Ethics. Major granting agencies have policies and guidelines that researchers must follow. Research or academic institutions all have their own review process for all projects involving humans or animals. Ethics Review Boards normally have forms to be completed describing the research project, the recruitment of volunteers, the inclusion and exclusion criteria, appropriate consent forms, and many other important questions regarding the storage of the data and guarantees of privacy and anonymity in reporting results. An ethics certificate is issued for each project with a timeline for which the certificate is valid.

Additional issues to be aware of are: conflicts of interest, plagiarism, protection and respect of intellectual property, confidentiality, privacy, and secrecy. The professional and positive interactions between engineers and physicians is important to ensure that research and development involving technology in a medical environment is successful (Bessenger, 1992; Daniels, 1992; Piehler, 1992). Another important consideration is the assessment of the impact of the planned work on society and on people. New developments, particularly in robotics, genetic engineering, and nanotechnology may raise serious concerns. The Center for the Study of Technology and Society summarizes the issue: “The same technologies that will let us cure diseases, expand the economy and overcome every day inconveniences can theoretically bring about catastrophes. Is the risk of an apocalypse serious enough for us to relinquish the current pace of technological innovation?”(Specht, 2000). Technology is often ahead of the development of societal guidelines and laws. Are we living in an era where technological determinism prevails? Can Society provide sufficient controls and moral and ethical guidance to prevent irreparable harm to our world and the people in it? Engineers must be able to understand how to verify the models that assess the impacts (positive and negative) of their technological development on society, and need a heightened awareness of universal responsibility and interdependability of engineers and society.

Ensuring that research and clinical services are equitable and inclusive of both sexes and various races represented in the population is also an important part of an ethical approach. Until recently, most clinical studies were performed exclusively with male subjects. New studies seem to be more inclusive of both sexes and of various races. However, gender inequalities still exist in all parts of the world, developing and developed, but in various degrees. This pertains not only to research, but to access to clean water, food, health care, education, and all aspects of life and security (Moreno, 2005, pp. 476–483). So in planning new studies, researchers and activists must assess how the planned work or project affects people (women, men, and children), and how the work can improve life on earth for everyone.

CONCLUSION

Performing research on humans or animals, or even on archived data or human material, must follow strict guidelines and protocols to ensure that all ethical principles are respected. This includes applying for an Ethics Certificate for each project, while ensuring that granting agencies’ policies are met, regarding good research practices, as well as all codes of ethics that apply to the particular situation. It may be appropriate to apply ethical theories to examine whether all aspects of the project are deemed ethical. It is also important to keep in mind that technological developments can bring major benefits to society, but may also bring hazards in their wake. Engineers should include in their plans an assessment of negative impacts on society, and attempt to minimize these. Engineers are also well placed to help solve some of the world’s largest problems and challenges as outlined in the United Nations Millennium Development Goals (United Nations, 2005). In the long run, society must institute laws and ethical codes of conduct to guide the direction and impacts of future developments. An overarching principle to keep in mind is the dynamic nature of ethical issues, so guidelines and principles must be reviewed regularly to assess their relevance.

REFERENCES

Andrews, G. C. (2005). *Canadian professional engineering and geoscience: Practice and ethics.* Toronto: Thomson Nelson.

Bessinger, C.D. Jr. (1992). Life-systems ethics and physician-engineer interactions. *Journal of Investigative Surgery, 5*(3), 185–190.

Daniels, A. U. (1992). Ethics education at the engineering/medicine interface. *Journal of Investigative Surgery, 5*(3), 209–218.

Frize, M. (1996). Teaching ethics & the governance of the profession: Ahead or behind professional practice realities? In *Proceedings of the 9th Canadian Conference on Engineering Education* (pp. 483–489).

Frize, M. (2005). The importance of teaching ethics to biomedical engineers. *International Journal of Medical Implants and Devices, 1*(2), 57–60.

Goodman, G. (1989). The profession of clinical engineering. *Journal of Clinical Engineering, 14*(1), 27–37.

Illingworth, P., & Parmet, W. E. (2006). *Ethical health care*. Upper Saddle River, N.J.: Pearson Prentice Hall.

Miles, S. H. (2004). *The Hippocratic oath and the ethics of medicine*. New York: Oxford University Press.

Moreno, J. D. (2005). *Is there an ethicist in the house? On the cutting edge of bioethics*. Bloomington and Indianapolis: Indiana University Press.

Piehler, H. R. (1992). Innovation and change in medical technology: Interactions between physicians and engineers. *Journal of Investigative Surgery, 5*(3), 179–184.

Quinn, M. (2005). *Ethics for the information age*. Boston: Pearson/Addison-Wesley.

Saha, S., & Saha, P. S. (1986). Ethical responsibilities of the clinical engineer. *Journal of Clinical Engineering, 11*(1),17–25.

Saha, S., & Saha, P. S. (1997). Biomedical ethics and the biomedical engineer: A review. *Critical Reviews in Biomedical Engineering, 25*(2), 163–201.

Specht, J. (2000). *Centre for the study of technology and society*. Retrieved February 5, 2008, from: http://www.tecsoc.org/innovate/focusbilljoy.htm

United Nations. (2005). *UN millennium development goals*. Retrieved February 5, 2008, from http://www.un.org/millenniumgoals/

World Medical Association. (2004). *World Medical Association Declaration of Helsinki*. Retrieved February 5, 2008, from http://www.wma.net/e/policy/b3.htm

Wueste, D. E. (1997). Professions, professional ethics, and bioengineering. *Critical Reviews in Biomedical Engineering, 25*(2), 127–149.

KEY TERMS

Act Utilitarianism (Act U): Theory stating that an action is good if it benefits someone, and bad if it harms someone, also known as the "greatest happiness principle" (individual moral actions).

Cultural Relativism (CR): Ethical theory where the meaning of right and wrong rests with a society's actual moral guidelines.

Declaration of Helsinki: Statement of ethical principles, developed by the World Medical Association, to provide guidance to physicians and others when performing research involving human subjects, identifiable human material, or data.

Divine Command Theory (DCT): Theory based on the idea that good actions are those aligned with the will of God; bad actions are those contrary to the will of God; that we owe obedience to our Creator, and that God is all-good and all-knowing, the ultimate authority.

Ethics: The philosophical study of morality, of right and wrong, of good and evil in human conduct.

Hippocratic Oath: Oath taken by physicians upon graduation, outlining the moral and ethical obligations to patients during the practice of medicine.

Kantianism: Theory also referred to as the categorical imperative, and pertaining to actions that are universally considered to be good, and involve good will and duty. Treat yourself and others as ends in themselves, never as a means to an end.

Morality: Defined as what people believe to be right and good, and the reasons for it.

Nuremberg Code: Code created as a result of the Nuremberg trials at the end of World War II, dealing with principles of human experimentation and voluntary informed consent.

Rawl's Theory of Justice: Theory stating that each individual may claim basic rights and liberties (ex., freedom of thought, speech, right to be safe from harm, and so on), as long as these claims are consistent with everyone having a claim to the same rights and liberties.

Rights Theory: Theory stating that everyone has rights arising simply from being born—the right to life, maximum individual liberty, and human dignity are all fundamental rights. This theory is the basis for the "Charter of Human Rights and Freedoms."

Rule Utilitarianism (Rule U): Theory stating that a moral rule should be followed, because its universal adoption would lead to the greatest increase in total happiness.

Social Contract Theory (SC): Theory based on the benefits to the community, governing how people are to treat one another. It is framed in the language of individual rights, and explains why rational people act out of self-interest in the absence of common agreement.

Subject Relativism (SR): Theory where there are no universal moral norms of right and wrong. Persons decide right and wrong for themselves.

European General Practitioners' Usage of E-Health Services

José Manuel Ortega Egea
University of Almería, Spain

María Victoria Román
University of Almería, Spain

Manuel Recio
University of Almería, Spain

INTRODUCTION

Technology plays an increasingly critical role in current markets. Diverse technological innovations are significantly changing how services are provided in different industries (Bitner & Brown, 2000). In this regard, new digital and Internet-related technologies are dramatically modifying how companies, employees, and customers interact in service encounters (Alba, Lynch, Weitz, Janiszewski., Lutz, Sawyer, & Wood, 1997; Hoffman & Novak, 1996).

In the health care sector, Internet technologies have the potential to improve the provision of health and patient care services like no other communications medium in the past (Shepherd & Fell, 1997). The health care sector, where interpersonal encounters between health staff and patients determine to a great extent service quality and patient satisfaction, will require diverse modifications, in order to fully benefit from the improvements promised by Internet technologies.

The term e-health is being widely used to refer to "health services and information delivered or enhanced through the Internet and related technologies" (Eysenbach, 2001). Physicians are largely turning to e-health services for gains in efficiency and improved quality of patient care. Factors related to higher constraints on physicians' costs, time, and private practices are currently accelerating Internet adoption among health professionals (Kassirer, 2001).

However, previous research shows that there are significant differences, both at the national and regional level, with regard to the development and adoption of Internet-related technologies (Hamill, 1997). These differences are even observable among developed countries in the European Union. In this regard, individuals from Northern European countries show higher rates of Internet adoption, as well as far more developed digital infrastructures. The unequal access to Internet technologies and slower Internet adoption processes in technologically backward countries ("digital divide") limit the potential positive effects of Internet technologies on the provision of health services in such regions.

Less clear is the influence of other external variables (e.g., gender and age) on physicians' professional use of Internet technologies. Diverse studies have dealt with this issue (Korgaonkar & Wolin, 1999), both in health and nonhealth contexts, with mixed results.

To effectively promote the use of Internet technologies for the provision of patient care services, it is important to understand the nature of physicians' current professional uses of the Internet. This article explores how the public Internet and dedicated networks are currently being used by European General Practitioners, and how external factors affect the degree of physicians' usage of e-health services.

BACKGROUND

Role of Technology and the Internet in Service Sectors

In the services sector, technological innovations have a great potential for improving the provision of diverse services. Parasuraman and Colby introduced the "services marketing pyramid model," which explicitly acknowledges the key role of technology in supporting and facilitating service delivery (Parasuraman & Colby, 2000).

The rapid impact of Internet technologies on service encounters has been documented in several studies (Parasuraman & Zinkhan, 2002). The use of the Internet by companies and consumers is radically changing how services are provided, both in B2C and B2B contexts. As a result, an increasing number of Internet-related studies is being carried out, focusing on the implications of the Internet in service industries (e.g., development of the e-SERVQUAL scale of service quality on the Web) (Zeithaml, Parasuraman, & Malhotra, 2000).

E-Health Research

A review of previous literature evidences that studies analyzing Internet adoption among health professionals are scarce (Sands, 2003). A recent report by the Boston Consulting Group offers one of the most detailed examinations of Internet usage by European physicians (Flanagan, Guy, Larsson, & Saussois, 2003). The main conclusions of this report include: (1) Most of physicians use the Internet for professional purposes, and it has an increasing impact on their professional knowledge, diagnoses, and prescriptions; (2) European physicians perceive the Internet as a very suitable medium for the acquisition of health information, continuous medical training, and complementary source of information for patients; (3) Physicians perceive that Electronic Health Care Records and Electronic Prescribing Systems contribute to an increased efficiency in their professional activities; (4) There is a trend towards more sophisticated Internet uses for increasing the quality of patient care services in the near future.

The review of previous studies evidences that further research is needed, especially empirical studies into the antecedents, current behavior, main purposes, and consequences of Internet usage for health purposes. The results of such studies should be taken into account by European, national, and regional authorities, in order to develop the suitable promotion strategies, aimed at further integrating new information technologies into patient care services.

RESEARCH OBJECTIVES

This research will draw on the main uses of Internet services and dedicated healthcare networks for physicians' professional purposes, in order to identify different segments of Internet users among European General Practitioners. Given the potential to improve doctor-patient relationships and health professionals' daily work, this article strives to clarify the extent to which general practitioners are taking advantage of Internet-related technologies in the European Union.

The following specific purposes have been set for this article:

1. Segmentation of European general practitioners, based on their adoption and actual professional uses of the Internet and General Practitioners networks. Behavioral variables have been selected for the characterization of segments.
2. Analysis of the effects of external variables (country, gender, age, location, and size of the medical practice) on the previously identified groups of Internet users.

The findings of the empirical analyses will clarify the existence and behavioral characteristics of the different segments (more vs. less advanced Internet users). Due to current national and regional differences in Internet adoption, these segments are expected to be differently distributed between European countries. The other covariates included in the study are not expected to exert significant effects on the identified segments.

METHODS

Sample

The empirical investigation is based on the survey "Flash-Eurobarometer Nº 126 – General Practitioners and the Internet" (2002), conducted on a sample of European general practitioners. The Eurobarometer surveys have been regularly carried out on behalf of the European Commission in all EU-countries. Table 1 shows the characteristics of the sample and the sampling procedure.

Statistical Analyses

The following methodology has been selected, according to the purposes of this study:

1. Latent Class Cluster Analysis has been applied, in order to identify homogeneous segments of Internet users among European physicians.

Table 1. Characteristics of the sample

SAMPLE	
Sampling procedure	• Telephone survey, using a "constant fraction" procedure, in order to guarantee the representativity of each national sample for its universe. The sampling has been carried out in each country on the basis of the published lists of General Practitioners.
Dates	• Interviews conducted between May 27th and June 19th, 2002
Sample size	• 3,512 respondents (aprox. 200 per country)
Collected information	• Sociodemographic variables: country, age, gender, location, and size of the medical practice. • Variables related to the use of diverse e-health services by European General Practitioners.

Latent Class Cluster Analysis overcomes some of the barriers involved in traditional clustering techniques. It is especially useful and flexible for dealing with categorical variables, which is the case in this study (Vermunt & Magidson, 2002).

2. The relative distribution of the identified segments with regard to several external variables has also been examined. Significance tests and measures of the association between the covariates and the segments are reported for this purpose.

RESULTS

Identification of Segments

From a total sample of 3,512 participants, the authors selected general practitioners who use Internet-enabling devices in their private practices (e.g., personal computers, laptops, or PDAs) for inclusion in subsequent statistical analyses. In most of doctors' private offices (2,862 or 81.49%) computer equipment is available. With regard to Internet use in the medical practice, a total of 2,264 professionals (64.46%) are currently accessing the Internet—or a General Practitioners network—from their private practices, compared to 96% of health professionals using Internet technologies the United States (Von Knoop, Lovich, Silverstein, & Tutty, 2003).

The Bayesian Information Criterion (BIC) was selected for the identification of the optimal number of segments. According to Biernacki and Govaert, in clustering contexts, the BIC criterion outperforms other classification criteria (Biernacki & Govaert, 1999). Taking into account the changes in BIC between the different cluster solutions (see Table 2), three is the optimal number of segments among European general

Table 2. BIC for different number of segment-solutions

	BIC (based on LL)	Change in BIC[a]
1 Cluster	81607,071	
2 Clusters	59574,165	-22032,906
3 Clusters-OPTIMAL	55693,742	-3880,423
4 Clusters	53417,221	-2276,521
5 Clusters	51889,497	-1527,724
6 Clusters	50682,053	-1207,444
a) Changes refer to the previous number of clusters in the table		

practitioners currently using the Internet or a general practitioners network. The classification error for the three-cluster solution is 2.36%.

Including the group of "Non-Internet Users," European physicians can be classified into four segments, according to their current professional use of the Internet or a General Practitioners network.

Behavioral Characterization of Segments

Segment 1: "Information Searchers" (28.99%)

- The use of Internet-based networks to search for health-related information is the main characteristic of this segment. Internet uses related to the acquisition of health information include: accessing medical associations' Web sites for professional-oriented information; searching for prescribing information; searching information online for the purpose of continuing education; and consulting medical journals.
- These doctors do not transfer patient-identifiable data through the Internet or a General Practitioners network.
- Moderate use of Electronic Health Care Records. EHCR systems are specifically purchased for the medical practice.
- These professionals do not use Web sites for their practices very often.

Segment 2: "Advanced Users" (20.96%)

- Like "Information Searchers," the search of health information for professional purposes through Internet networks is a quite-relevant Internet use for general practitioners of this group.
- The key distinguishing characteristic of this segment relates to the use of the Internet-based networks for transferring patient-identifiable data. Practically all members of this segment send or receive this kind of patient data online. The main purposes of these data transfers include: submitting patient care reimbursement claims; receiving results from laboratories and other diagnostic procedures; transferring administrative patient data to a secondary care provider; and transferring patient medical data to other medical care providers for the purposes of continuity of care.
- E-mail communication with patients is not a widespread activity yet, even for this group of "Advanced Users".
- Internet technologies are rarely used to offer telemedicine or electronic prescribing services to patients.
- The majority of "Advanced Users" does not obtain specific patients' consent, or use electronic signatures in communicating patient medical data online.
- The use of Electronic Health Care Records is moderate.

Segment 3: "Laggards" (14.52%)

This segment shows a less sophisticated use of Internet-based networks for professional purposes in all analyzed areas: limited search for medical information online; these physicians do not transfer patient-identifiable data through the Internet or a general practitioners network; and these physicians rely less frequently on EHCR and Web sites for the private practices.

Segment 4: "Non-Internet Users" (35.54%)

It is important to emphasize the high share of European health professionals that are currently not taking advantage of the new possibilities afforded by Internet technologies for the provision of health care services.

Characterization of Segments According to External Variables

An attempt has been made to characterize the identified segments, according to several external variables. These analyses should contribute to the debate on whether geographic criteria, demographic traits, or job characteristics show significant effects on the acceptance and use of Internet technologies and other technological innovations.

Country

One of the main purposes of this study relates to the examination of potential national differences, related to European physicians' use of the Internet or a general practitioners network. Based on the results of the LC Cluster Analysis, the effects of the "country" variable on the identified groups of Internet users is very significant (Wald=467.5118; P<.001), which suggests the existence of significant national differences between European countries.

It has also been investigated if the four identified segments, including the Non-Internet Users group, show statistical significant differences between countries. The relative distribution of the identified segments in the analyzed countries has been displayed in a contingency table. Again, the results of the chi-square test (chi-square=1,515.977; P<.001) suggest the existence of significant differences between countries.

Nevertheless, in large samples sizes, relationships can be found significant even when the association of variables is very weak. In these cases, measures such as Phi, Cramer's V, or the Contingency Coefficient should be reported, in order to fully represent the real strength of the association between the segments and the covariates. Phi, Cramer's V, and the Contingency Coefficient measure the strength of the relationship between two nominal variables in a contingency table, with values between 0 (no relationship) and 1 (perfect relationship). Phi should be used in 2x2 tables, while Cramer's V or the Contingency Coefficient should be reported when at least one of the variables has more than two categories.

Due to the large sample size in this study (3,512), the authors also provide measures of association between the segments and the covariates. In the case of the "country" variable, the values of Cramer's V and the Contingency Coefficient are 0.379 and 0.549, respectively. These values confirm the existence of strong country influences on the segments.

The following national trends in the use of the Internet and dedicated healthcare networks among European general practitioners can be observed, which can be examined in graphical form in Figure 1:

1. **Laggard countries:** The number of "Non-Internet Users" is very high in these countries. Relatively few "Advanced Users" and "Information Searchers" can be found. Countries included in this group are Greece, Spain, Germany, and Portugal.
2. **Follower countries:** The number of "Advanced Users" and "Information Searchers" is moderate, but the number of "Non-Internet Users" is still quite high. The following countries can be included in this group: Belgium, France, Ireland, Italy, Luxemburg, Holland, and Austria.
3. **Advanced countries:** These countries are characterized by very few "Non-Internet Users." There is a high number of "Advanced Users" and "Information Searchers." Advanced countries include Sweden, the United Kingdom, Finland, and Denmark. Denmark stands out as the country with the highest number of "Advanced Users."

While country differences in the adoption of e-health services resemble, to a great extent, traditional economical and infrastructural differences between European countries, attitudinal, perceptual, or cultural factors also exert significant influences on the acceptance of Internet technologies for health professional purposes. Further research should examine the role and relative importance of physicians' individual characteristics in the adoption of e-health services.

Gender

Based on the results of the LC Cluster Analyses, the effects of "gender" on the identified groups of Internet users are significant (Wald=17.8737; P<.001).

The tests of gender differences between the four segments provided the following results (chi-square=17.821; P<.001; Cramer's V=0.071; Contingency Coefficient=0.071). As suggested by these measures, there is a very weak association between gender of surveyed physicians and their classification in each of the identified segments.

These results show very similar profiles for men and women with regard to the classification of physicians in the different segments. Although the distribution patterns of the identified "clusters" among men and women are very similar, it can be observed that male physicians show a slightly more sophisticated use of Internet technologies for their medical practice. Among users of the Internet or a general practitioners network, there is a higher percentage of women included in the "Laggards" group (18.39%), compared to 13.19% of men. For men, the higher number of "Advanced Users" and "Information Searchers" balances the relative less importance of the "Laggard" group.

As a conclusion, these results confirm previous research, suggesting that gender is becoming less significant in explaining technology usage, especially in highly educated groups, like physicians.

Age

This study has also examined if there are significant age differences in physicians' adoption of Internet-related technologies for professional purposes. According to the results of the LC Cluster Analysis, the effects of "age" on the identified classes of Internet users are statistically significant (Wald=39.8204; P<.001).

Further analysis of age differences between the four segments showed the following results: chi-square=255.258; P<.001; Cramer's V=0.157; Contingency Coefficient=0.262. These measures indicate a moderate association between physicians' age and their classification in one of the identified segments.

There seems to be a clear trend, with regard to the relationship between age and the use of the Internet or a general practitioners network (see Figure 2). Younger physicians (less than 35 years old) show the higher percentage of Internet users. Physicians in ages between 36 and 40 years present the most sophisticated use of e-health technologies (highest number of "Advanced Users" and "Information Searchers"). In general, physicians between 36 and 50 years old show quite similar uses of Internet technologies. The results point to a clear declining trend relating age and the use of Internet-based networks for physicians in ages over 50 years. Physicians between 56 and 60 years old show a very sophisticated use of e-health technologies, but a comparatively less total number of Internet users. Adoption rates of e-health technologies among physicians over 60 years old are very limited.

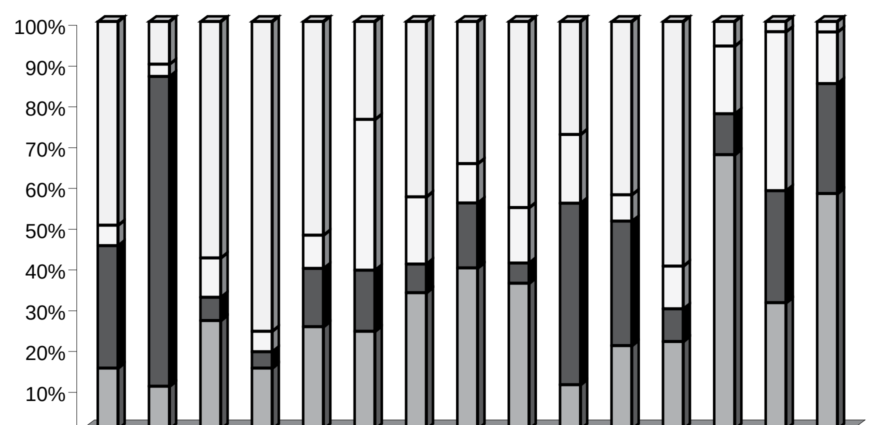
Location of Medical Practice

The Wald statistic and the z-values for the external variable "location of practice" indicate that this covariate does not exert a statistically significant effect on the identified segments of Internet users (Wald=1.8683; *P*=0.76).

Further tests confirm the existence of very homogeneous usage patterns of technological healthcare networks for physicians working in different settings, whether metropolitan zones, other town, or rural areas (chi-square=24.663; *P*=0.03; Cramer's V=0..048; Contingency Coefficient=0..084).

Although not statistically significant, it is worthwhile to note that general practitioners working in rural settings show more sophisticated usage patterns of e-health technologies than physicians in metropolitan zones, with a higher share of "Advanced Users" and "Information Searchers." Thus, possible concerns about the existence and widening of a "digital divide" between metropolitan and rural areas (Brodie, Flournoy, Altman, & Blendon, 2000) have not been confirmed for European health practitioners.

Figure 1. Distribution of segments in EU-countries

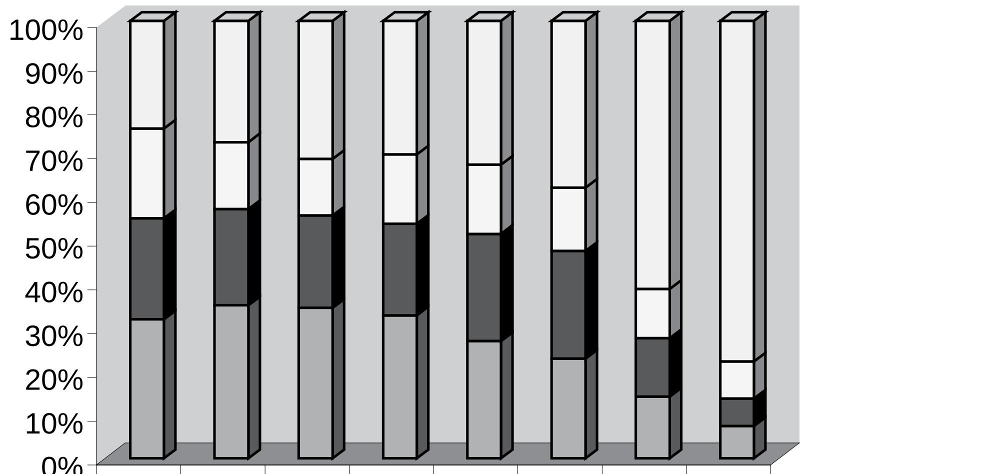


	Belgium	Denmark	Germany	Greece	Spain	France	Irland	Italy	Luxembu rg	Holland	Austria	Portugal	Finland	Sweden	United Kingdom
Non-Internet Users	50,00%	10,50%	58,00%	76,00%	52,43%	24,00%	43,00%	34,88%	45,68%	27,72%	42,50%	60,00%	6,00%	2,50%	2,60%
Laggards	5,00%	3,00%	9,67%	5,00%	8,11%	37,00%	16,50%	9,63%	13,58%	16,83%	6,50%	10,50%	16,67%	39,00%	12,66%
Advanced Users	30,00%	76,00%	5,67%	4,00%	14,32%	15,00%	7,00%	15,95%	4,94%	44,55%	30,50%	8,00%	10,00%	27,50%	26,95%
Information Searchers	15,00%	10,50%	26,67%	15,00%	25,14%	24,00%	33,50%	39,53%	35,80%	10,89%	20,50%	21,50%	67,33%	31,00%	57,79%

*The exhibit shows the relative importance of the identified segments, separately for each country.

Information Searchers Advanced Users Laggards Non-Internet Users

Figure 2. Age differences



	<35	36-40	41-45	46-50	51-55	56-60	61-65	>66
Non-Internet Users	24,62%	27,72%	31,57%	30,53%	32,84%	38,15%	61,27%	77,89%
Laggards	20,51%	15,28%	12,94%	15,87%	15,85%	14,46%	11,27%	8,42%
Advanced Users	23,08%	22,02%	21,09%	20,91%	24,51%	24,62%	13,38%	6,32%
Information Searchers	31,79%	34,97%	34,41%	32,69%	26,80%	22,77%	14,08%	7,37%

*The exhibit shows the relative importance of the identified segments, separately for each age-group.

Size of Medical Practice

The performed LC Cluster Analysis indicates that the effects of "size of practice" on the identified groups of Internet users are not statistically significant (Wald=6.4271; *P*=0.38).

After the inclusion of "Non-Internet Users" group, further tests suggest the existence of a statistically significant relationship between this covariate and the identified segments (chi-square=326.744; *P*<.001; Cramer's V=0..176; Contingency Coefficient=0..292). The value of Cramer's V points to a weak increasing relationship between size of the medical practice and the usage rates of Internet-based healthcare networks. This association should be interpreted with caution, as the results of the Latent Class Cluster Analysis were not statistically significant.

FUTURE TRENDS

Future research should further clarify how the adoption of Internet-based networks is taking place in the health care sector. The use of perceptual and attitudinal measures towards the Internet, and new digital technologies would offer a richer explanation of the relative role played by physicians' psychological characteristics in the adoption processes of Internet-based networks.

The acceptance of e-health services among health professionals other than general practitioners should also be researched. A sample including surgeons, specialists, nurses, and other professionals would clarify if the acceptance of the Internet or dedicated healthcare networks is affected by the professional category of health staff.

CONCLUSION

Segmentation of European Physicians

This research has identified four segments of European physicians, based on their actual uses of Internet technologies for health-related purposes. It has also been explored how external variables are associated with the use of e-health services by general practitioners.

The identified groups have been labeled as "Information Searchers," "Advanced Users," "Laggards," and "Non-Internet Users," according to the degree and purposes of physicians' health-related Internet usage.

The key distinguishing difference between "Advanced Users" and "Information Searcher" relates to the use of the Internet-based networks by "Advanced Users" for transferring patient-identifiable data. Other behavioral differences between these groups can be observed, related to other professional uses of the Internet or a general practitioners network, but not as significant.

The identification of these segments should provide a valuable basis for the development of suitable Internet promotion strategies, adapted to the needs of the different groups. Strategies targeted at the "Laggards" and "Non-Internet Users" groups should focus on explaining the benefits provided by Internet technologies in general, not only for health purposes. Training should be offered to these groups, where the needed skills for Internet use can be acquired. At later stages, these people should be offered continuous education on the possibilities afforded by Internet technologies for their professional activities. Due to usual time and costs' constraints on physicians, these courses should emphasize the potential benefits of Internet-technologies in terms of efficiency and improved patient satisfaction.

Policies targeted at "Information Searchers" should focus on explaining the potential benefits of using the Internet or a dedicated healthcare network for transfers of patient data. Both "Information Searchers" and "Advanced Users" receive training on how e-mail patient communication, electronic prescribing, and Web sites can help them to improve the management of their professional practices, as these are not widespread activities yet.

Advanced Users' main characteristic relates to the use of Internet technologies for transfers of patient-identifiable data. Nevertheless, the protection of patients' privacy, such as obtaining patients' consent or the use of electronic signatures, seems to be less than desirable. For this purpose, a homogeneous European legal framework should provide clear guidelines on this matter and ensure the protection of patients' privacy. More sophisticated e-health uses, such as telemedicine services, should be continually introduced to more advanced users at later stages.

Country Differences

With regard to the existence of national or regional differences in the use of Internet-based healthcare networks between European physicians, Northern European countries (Sweden, Finland, Denmark, and the United Kingdom) are clearly on the forefront, while Internet adoption by physicians from Southern countries (Greece, Spain, and Portugal) is clearly developing at slower paces.

Although certain health-related Internet uses, such as transfers of patient data online, can contribute to substantial improvements in the provision of patient care services (continuous patient monitoring, communication with patients, contact with dispensing pharmacists, and so on), the differences found between countries can limit the potential benefits of these Internet-related uses. Concerns about patients' privacy protection may slow down the adoption process of such Internet-based services in the health care sector, and different national regulations in this matter may contribute to widen regional disparities in the European Union. In countries like Spain, e-health services are subject to governmental regulations. Thus, generalization of certain services, such as electronic prescribing with full legal protection, is becoming a rather slow process.

It is also worth considering that the "North/South divide" in Internet adoption is also evident, not only among health professionals, but also with regard to the general public. If the potential of Internet technologies for improving the provision of patient care services is to be fully realized on a European basis, European and national authorities should play a key supporting role.

In Europe, the commitment of national and regional authorities with the policies developed by the European Commission will be critical for their success. The eEurope Action Plan 2005, with regard to the promotion of e-health services, aims to build a homogeneous Europe-wide legal framework, which ensures the secure transfer of patient-identifiable data over the Internet and dedicated healthcare networks.

Governmental support should mainly focus on promoting the adoption and use of Internet technologies (e.g., through marketing campaigns and training), developing modern digital infrastructures, and reducing negative attitudes of physicians and patients towards the use of the Internet for health-related purposes (e.g., privacy and security concerns).

Differences in Gender, Age, Location, and Size of Practice

Previous research has provided mixed results on whether external variables such as demographics and geographic criteria are significant predictors of Internet and technology usage. The relative influence of several *covariates* on physicians' use of e-health services has been explored in this study, in addition to the existence of country influences. In general, these external variables have provided rather weak relationships with the identified segments in this research. Of the analyzed variables, only a physicians' age shows a moderate decreasing association with the use of the Internet or a general practitioners network.

Gender, location, and size of the medical practice have not provided significant effects on physicians' use of e-health services. The nonexistence of significant differences in the use of Internet-related technologies for health purposes, based on "location of the medical practice," has important implications, as e-health services have a great potential for improving the protection of patients' health in rural and isolated areas. The prerequisite of the acceptance of Internet-based technologies by health care providers in rural settings seems to have been achieved. In order to ensure the success of e-health services in nonmetropolitan areas, European, national, and regional authorities should further promote Internet use by citizens living in smaller towns and rural areas.

REFERENCES

Alba, J., Lynch, J., Weitz, B., Janiszewski., C., Lutz, R., Sawyer, A., & Wood, S. (1997). Interactive home shopping: Consumer, retailer, and manufacturer incentives to participate in electronic marketplaces. *Journal of Marketing, 61*(3), 38–53.

Biernacki, C., & Govaert, G. (1999). Choosing models in model-based clustering and discriminant analysis. *Journal of Statistical Computation and Simulation, 64*, 49–71.

Bitner, M.J., & Brown, M.L. (2000). Technology infusion in service encounters. *Journal of the Academy of Marketing Science, 28*(1), 138–149.

Brodie, M., Flournoy, R.E., Altman, D.E., & Blendon, R.J. (2000). Health information, the Internet, and the digital divide. *Health Affairs, 19*(6), 255–265.

Eysenbach, G. (2001). What is e-Health? *Journal of Medical Internet Research, 3*(2), e20. Retrieved January 31, 2008, from http://www.jmir.org/2001/2/e20/

Flanagan, A., Guy, P., Larsson, S., & Saussois, C. (2003). European physicians and the Internet. *The Boston Consulting Group Report*. Retrieved January 31, 2008, from http://www.bcg.com/publications/files/Euro_Physicians_Internet_Rpt_Mar03.pdf

Hamill, J. (1997). The Internet and international marketing. *International Marketing Review, 14*(5), 300–323.

Hoffman, D.L., & Novak, T.P. (1996). Marketing in hypermedia computer-mediated environments: Conceptual foundations. *Journal of Marketing, 60*(3), 50–68.

Kassirer, J.P. (2001). Patients, physicians, and the Internet. *Health Affairs, 19*(6), 115–123.

Korgaonkar, P.K., & Wolin, L.D. (1999). A multivariate analysis of Web usage. *Journal of Advertising Research, 39*(2), 53–68.

Parasuraman, A., & Colby, C.L. (2000). *Techno-ready marketing: How and why your customers adopt technology.* New York: The Free Press.

Parasuraman, A., & Zinkhan, G.M. (2002). Marketing to and serving customers through the Internet: An overview and research agenda. *Journal of the Academy of Marketing Science, 30*(4), 286–295.

Sands, D.Z. (2003). Electronic patient-centered communication: Managing risks, managing opportunities. *The American Journal of Managed Care, 5*(12), 1569–1571.

Shepherd, C.D., & Fell, D. (1997). Health care marketing and the Internet. *Marketing Health Services, 17*(3), 50–51.

Vermunt, J.K., & Magidson, J. (2002). Latent class cluster analysis. In J.A. Hagenaars & A.L. McCutcheon (Eds.), *Applied Latent Class Analysis* (pp. 89–106). Cambridge, UK: Cambridge University Press.

Von Knoop, C., Lovich, D., Silverstein, M.B., & Tutty, M. (2003). Vital signs: E-health in the United States. *The Boston Consulting Group Report*. Retrieved January 31, 2008, from http://www.bcg.com/publications/files/Vital_Signs_Rpt_Jan03.pdf

Zeithaml, V.A., Parasuraman, A., & Malhotra, A. (2000). *A conceptual framework for understanding e-service quality: Implications for future research* *and managerial practice*. Marketing Science Institute Working Paper. Retrieved January 31, 2008, from http://www.msi.org/msi/publication_summary.cfm?publication=17

KEY TERMS

Dedicated Networks: Digital networks dedicated to healthcare services.

Digital Divide: Gap between individuals and communities with greater and lesser access to digital technologies.

E-Health: Health services and information delivered or enhanced through the Internet and related technologies.

Electronic Health Care Records: Digital medical records which contain all relevant information of a person's health.

Electronic Prescribing: Use of digital networks and devices to automate the drug prescription process.

General Practitioners: Physicians whose practice is not oriented to a specific medical specialty.

LC Cluster Analysis: Latent class cluster analysis, which represents a robust and flexible statistical methodology for segmentation studies.

Segmentation: Process of identifying subgroups of people with similar behavioral or attitudinal patterns.

Telemedicine: Use of information technologies for the delivery of clinical care (e.g., home monitoring of patients through the Internet).

Evaluation of External Cardiac Massage Performance During Hypogravity Simulation

G. Dalmarco
Pontifical Catholic University, Brazil

T. Russomano
Pontifical Catholic University, Brazil

A. Calder
University of Glasgow, Scotland

Felipe Falcão
Pontifical Catholic University, Brazil

D. F. G de Azevedo
Pontifical Catholic University, Brazil

Subhajit Sarkar
University of Birmingham, UK

Simon Evetts
Thames Valley University, UK

Samuel Moniz
Universidade do Porto, Portugal

INTRODUCTION

Preservation of astronaut crew health during an exploration mission to the Moon or Mars will be crucial to mission success. The likelihood of a life-threatening medical condition occuring during a mission to Mars has been estimated by NASA to be 1% per year (Johnston, 1998; Johnston, Campbell, Billica, & Gilmore, 2004). Since basic life support is a vital skill in critical care medicine, plans must be in place for cardiopulmonary resuscitation in both microgravity and hypogravity (i.e., on the surface of the Moon or Mars).

Following the design of a body suspension device to simulate a hypogravity environment, subjects performed external chest compressions in 1G, 0.17G (Lunar), 0.38G (Mars), and 0.7G (Planet X). Chest compression adequacy was assessed by means of rate and depth. Heart rate immediately before and after three minutes of chest compression gave a measure of rescuer fatigue. Elbow flexion was measured using an electrogoniometer in order to assess the use of arm muscles to achieve chest compressions.

This study found that the mean depth (Lunar and Mars) and rate (Mars) of chest compression was below American Heart Association recommendations during hypogravity simulation in the female group. Furthermore, elbow flexion proved to be significantly greater during Lunar and Mars hypogravity simulation than that of the 1G control condition, suggesting that upper arm force may be used to counter the loss of body weight in an attempt to maintain adequate chest compression under these conditions.

BACKGROUND

A new initiative announced by the United States aims to resume manned lunar expeditions with a landing planned for 2015, and ultimately a manned mission to Mars (White House Press Secretary, 2004). Japan's long-term vision resembles those of U.S. President George W. Bush and European space officials, who hope to land astronauts and robots on the Moon as a first step to sending a manned space mission to Mars

(Exploration of Mars and the Moon, 2006; Future of Human Space Flight, 2006). Over the next decade, the Japan Aerospace Exploration Agency (JAXA) plans to call for scientists to develop robots and nanotechnology for surveys of the Moon and design a rocket and space vessel capable of carrying cargo and passengers. By 2015, JAXA will review whether it is ready to pour resources into manned space travel, possibly building a base on the Moon. A decision to try for Mars and other planets would be made after 2025 (Japan Space Flight History, 2005). Plans regarding manned trips to the Moon and Mars are not restricted to the United States and Japan. European and Russian scientists have set out a route map for manned missions to Mars that aims to land astronauts on the Red Planet in less than 30 years. A human mission to the Moon, proposed for 2024, would demonstrate key life-support and habitation technologies, as well as aspects of crew performance and adaptation to long-distance space flight. The ESA has planned two flagship missions to Mars-ExoMars that would land a rover on the planet in 2009, and Mars Sample Return would bring back a sample of the Martian surface in 2011-2014. Other test missions will include an unmanned version of the flight that would eventually carry astronauts to Mars (ESA, 2006).

The diagnosis and treatment of acute and chronic medical conditions have been identified by all space agencies as one of the highest priorities for both current orbital space flight and future exploration class missions to the Moon and Mars. In particular, there is a need for evidence-based guidelines and techniques for management of medical emergencies during such missions. To date, all mortalities in space flight have been the consequence of sudden catastrophic technological failures, which have left no opportunity for corrective action (Telemedical Emergency Management, 2004). However, significant morbidity has occurred, including type I decompression sickness (Apollo 11), urinary tract infection (Apollo 13), cardiac arrhythmias (Apollo 15, Mir station 1987), N_2O_4 pneumonitis (Apollo-Soyuz test project), and prostatitis, sepsis, and hypothermia (Salyut 7). These health issues have thus far not required advanced life support measures but have on occasion required the prophylactic deorbiting of the crew member (Telemedical Emergency Management, 2004).

A higher incidence of acute and life-threatening medical problems may be anticipated with the increasingly longer space missions and extravehicular activity requiring increased physical labor as developmental activity in space progresses. In addition, selection criteria in terms of age and health are in many ways less stringent than those of the early days of human space flight. Thus the rare but potential likelihood of serious medical incidents, such as a life-threatening cardiac event, must be anticipated and adequately prepared for.

Cardiac arrest can be in the form of asystole, pulseless ventricular fibrillation or electromechanical dissociation, which are the results of primary cardiac diseases such as coronary artery disease or a cardiomyopathy, trauma, or secondary to diseases of other systems (e.g., respiratory failure). Since its introduction in the 1960s (Safar, 1961), modern cardiopulmonary resuscitation (CPR) with manual chest compressions at its foundation has been shown to improve survival after cardiac arrest (Cummins, Eisenberg, Hallstrom & Litwin, 1985; Ritter et al., 1985). Many studies have confirmed the haemodynamic significance of delivering consistent, high quality, infrequently interrupted chest compressions (Wiggington, Miller, Benitez & Pepe, 2005).

Basic life support and CPR are essential features of current astronaut training. NASA astronauts are currently trained in two techniques. The first utilizes the Crew Medical Restraint System on the International Space Station to restrain both patient and provider so that chest compressions are performed in a manner similar to the standard terrestrial method. The second is the so-called Hand Stand method (Hamilton, 2003; Jay, Lee, Goldsmith, Battat, Jaurrer & Suner, 2003). These and other available options, such as the Evetts-Russomano technique (Evetts, 2004; Evetts, Evetts, Russomano, Castro, & Ernsting, 2004; Evetts, Evetts, Russomano, Castro, & Ernsting, 2005) developed by the King's College London, UK, and the Microgravity Laboratory PUCRS, Brazil, have been based on previous and ongoing research to address the need to perform CPR in orbital microgravity. In this environment, the delivery of rescue breaths does not present a major technical problem. The main challenge is the application of sufficient force to the victim's chest in the absence of weight, which provides the main source of compression force under the standard terrestrial method. In the absence of data to suggest alternative requirements, all the above methods have aimed to match the terrestrial standards of CPR set by the American Heart Association (2006) and the European Resuscitation Council (Handley, Monsierus, & Bossaert, 2001).

Hypogravity of various magnitudes will be the condition experienced during the planetary or lunar surface elements of future exploration missions. CPR methods developed for microgravity may not be applicable during the surface elements of these missions, as gravity is present, albeit reduced. In addition, the standard terrestrial method, which involves the application of the weight of the provider as the compression force delivered to the sternum, may also not be applicable if weight is reduced to an extent that the force achieved no longer produces adequate chest compressions. Being able to achieve adequate depth is crucial to the effectiveness of external cardiac massage. Since optimally performed CPR provides at best only a third of normal cardiac output (Paradis et al., 1989), even modest deteriorations in performance of the provider may have significant effects on reducing effectiveness (Ashton, McCluskey, Gwinnutt & Keenan, 2002). The American Heart Association guidelines recommend a chest compression depth of 40-50mm at a rate of 100 compressions per minute as adequate, which is currently applied to nonterrestrial cardiopulmonary resuscitation (CPR).

Given the high priority of basic life support in astronaut training, it is important to develop an effective CPR technique for these altered gravitational environments. A previous theoretical study has suggested that if terrestrial standards of chest compression depth were to be adopted, they were unlikely to be efficient in Lunar or Martian gravities. This can only be overcome if the provider is of high mass and the patient's chest of high compliance (Sarkar, 2004).

MAIN FOCUS OF THE CHAPTER

A pioneer study was conducted to evaluate the performance of external chest compressions during hypogravity simulation of different magnitudes at the Microgravity Laboratory/PUCRS.

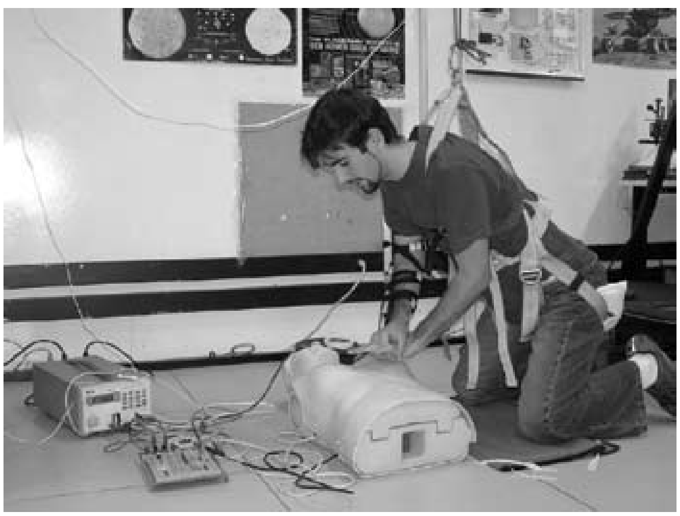
Three gravitational environments were simulated: 0.17G (Lunar), 0.38G (Martian), and 0.70G (Planet X). Planet X was created in order to provide an intermediate gravitational field to help with data analysis.

External chest compression adequacy was studied, along with arm flexion (upper-limb movement) during compressions. This research provides insight into cardiac massage performance in hypogravity and helps to reflect on performance in clinical and training environments.

Details of the equipment developed for this study are presented next.

Body suspension device. A method of simulating a reduced gravitational field has been developed by the Microgravity Laboratory/PUCRS. The body suspension system comprises a body harness, counterweights, and a load cell. The structure is pyramidal and consists of steel bars of 6cm x 3cm thickness. It has a rectangular base area of 300cm x 226cm and a height of 200cm. A steel cable connects the counterweights through a

Figure 1. Male subject performing external chest compression wearing the harness for hypogravity simulation and an electrogoniometer attached to his right arm



system of pulleys to a harness worn by the subject (Figure 1).

It uses a counterweight system consisting of 20 steel bars of 5kg each, placed opposite the subject. The necessary counterweights were calculated as follows: using equation 1, the relative mass of a subject in a simulated gravitational field can be calculated, where RM = relative mass (kg), BM = body mass on Earth (kg), SGF = simulated gravitational force (m/s^2), and 1G = $9.81 m/s^2$. Equation 2 gives the counterweight (CW, in kg) necessary to simulate body mass in a preset hypogravity level.

$$RM = \frac{0.6BM \times SGF}{1G} \quad (1)$$

$$CW = 0.6BM - RM \quad (2)$$

Basic life support training manikin. A standard full-body CPR mannequin (Resusci Anne SkillReporter, Laerdal Medical Ltd, Orpington, UK) was instrumented for measurement of external chest compression depth and rate by the Microgravity Laboratory, PUCRS. A 10KΩ potentiometer was used for chest compression depth measurement, which gave a real-time feedback to the subject by means of a LM3941 component that shows a voltage variation by lighting a line of colored LEDs (light emitting diodes) for depths of 0-28mm (red), 29-39mm (yellow), 40–50mm (green), and 51-60mm (red) (Figure 2). Chest compression frequency was indicated by an electronic metronome set to a frequency of 100/minute.

Load cell. A load cell was attached to the top of the pulley system in order to measure the corrected weight. The load cell consists of an Al 6351 aluminium tube (inner radius of 18mm, external radius of 32mm, and depth of 22mm). The inner part of the tube contains a Wheatstone bridge made with strain gauges, which vary resistance according to either compression or traction of the load cell. Calibration tests were performed to evaluate the linearity of the load cell. Two different weight loads (2kg and 17.45kg) were used for calibration before the beginning of each experiment.

Elbow electrogoniometer. The angle of elbow flexion (upper-limb movement) was measured using an electrogoniometer, which consists of two aluminium bars (20cm x 2.0cm x 0.3cm) covered with rubber material and connected by a 10KΩ potentiometer (Figure 3). A 5V power source was used. Calibration was performed before the beginning of each experiment by attaching a manual goniometer to the electrogoniometer. The full range voltage at 0º and 90º was then recorded.

Data acquisition system. A DataQ acquisition device (DATA-Q Instruments Inc, Akron, OH, USA) with 8 analogue and 6 digital channels, 10 bits of measurement accuracy, rates up to 14400 samples per second and USB interface was used. It supports full-scale range of ±10V and a resolution of ±19.5mV. WinDaq data acquisition software allowing for conversion of volts to the necessary unit was used. Three channels

Figure 2. External chest compression depth indicator and metronome

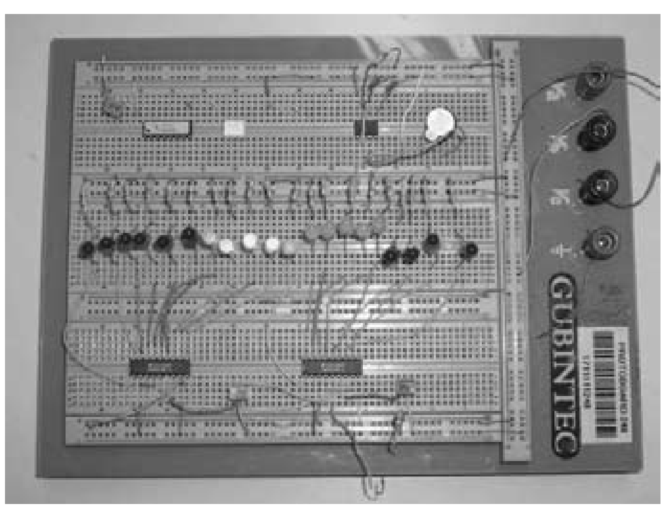


Figure 3. Elbow electrogoniometer

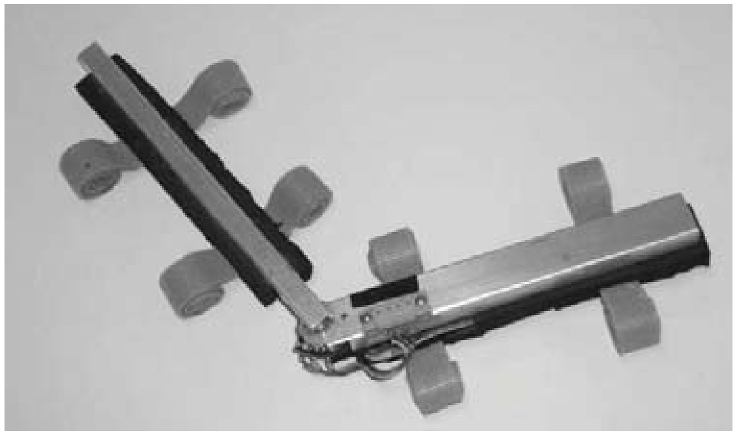
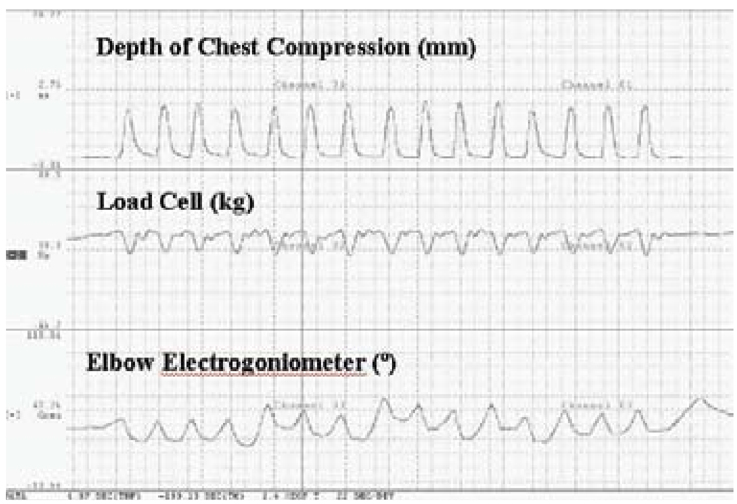


Figure 4. WindaQ during data collection



were used during data collection: depth and rate of chest compression, the amplified Wheatstone bridge signal of the load cell, and the elbow electrogoniometer signal (Figure 4).

External cardiac compression measurements were obtained from 10 men and 10 women. Subject characteristics are outlined in Table 1. Subjects followed a protocol of 15 chest compressions alternating with 10 seconds of rest for a total of three minutes. The 10-second rest simulated the time that would be taken for the provision of two mouth-to-mouth ventilations.

A five-minute training session was carried out by all subjects. On different days, a control phase of three minutes at 1G was followed by three minutes of one of three randomized levels of hypogravity simulation: 0.17G (Lunar gravity), 0.38G (Mars gravity), and 0.7G (Planet X).

During each visit, subject weight and height were measured prior to instrumentation with the body suspension harness and elbow electrogoniometer. Heart rate was measured before and immediately after the completion of the chest compressions. Adequacy of external chest compression was assessed by means of chest compression rate and depth.

The study protocol was approved by the PUCRS Research Ethics and Scientific Committees. Each subject provided written informed consent before participating in the experiment.

Statistical analysis was performed using student t-tests and two-way ANOVAs assuming a level of significance of 5%.

Subject characteristics are shown in Table 1. Subjects were matched for age and body mass index.

In all subjects, heart rate increased following three minutes of external chest compression. The magnitude of this increase in heart rate tended to be greater for the female group (Table 2).

The rate of chest compressions was not different between 1G and hypogravity in the male group (Figure 5, Table 3). However, there was a trend toward chest compression rate being slower in the female group in hypogravity as compared with 1G (Figure 5, Table 4). Similarly, male subjects maintained an adequate chest compression depth in the three hypogravity environments. Depth of compression, however, was inadequate in the female group during Lunar and Martian gravitational environments as compared to 1G (Figure 6, Table 4). During hypogravity simulation, there was a significant difference in all groups at all hypogravity values in elbow flexion (Tables 3 and 4).

FUTURE TRENDS

Previous studies on CPR administered under terrestrial conditions have highlighted features of an individual that may affect their ability to deliver effective cardiac massage. These factors include mass, height, gender, and physical fitness of the CPR provider (Ashton et al., 2002; Larsen, Perrin & Galletly, 2002; Lucia et al., 1999). In a study conducted by Ashton et al. (2002) with 40 individuals (20 males and 20 females), rescuer fatigue was investigated during two consecutive three-minute intervals of continuous chest compressions separated by a 30-second break representing a pulse and ECG check. The CPR was to be performed on a Ressusci Anne® manikin at a rate of 100/min with a depth of 40mm to 50mm. The male subjects were significantly heavier and taller than their female counterparts. During the study, the rate of chest compressions were maintained at approximately 100/min by the rescuers; however, the number of satisfactory chest compressions decreased progressively during resuscitation ($p<0.001$) due to fatigue. Female subjects were found to achieve fewer

Table 1. Subject characteristics

	Male Mean (± SD)	Female Mean (± SD)	p value
Age (yr)	24 (1.64)	26 (5.65)	0.162
Height (m)	1.81(0.03)	1.65 (0.07)	0.000
Weight (kg)	74.2 (10.58)	57.1 (6.53)	0.000
BMI (Kg/mm)	23 (3.09)	21 (3.04)	0.139

BMI: body mass index

Table 2. Subject heart rate responses to three minutes of external chest compression

	Male	Female
	Mean (±) SD	Mean (±) SD
Resting HR (bpm)	83 ± 5.49	93 ± 8.94
Planet X Final HR	99 ± 17.58	124 ± 12.40
P value	0.018	0.000
Mars Final HR	102 ± 13.11	123 ± 8.62
P value	0.002	0.000
Moon Final HR	103 ± 21.61	129 ± 8.49
P value	0.009	0.000

HR = heart rate; SD = standard deviation; P value represents the difference between resting heart rate and final heart rate at each hypogravity level.

Figure 5. Chest compression frequency in four gravitational environments. Values along y axis represent compression rate (compressions/minute)

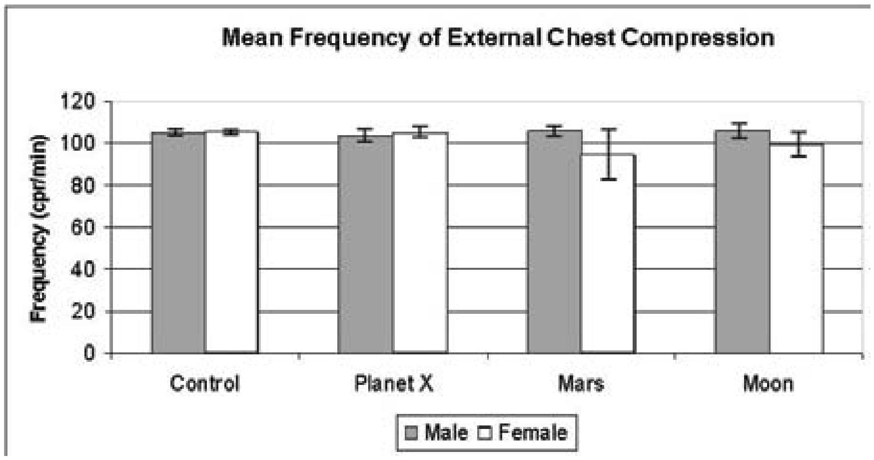
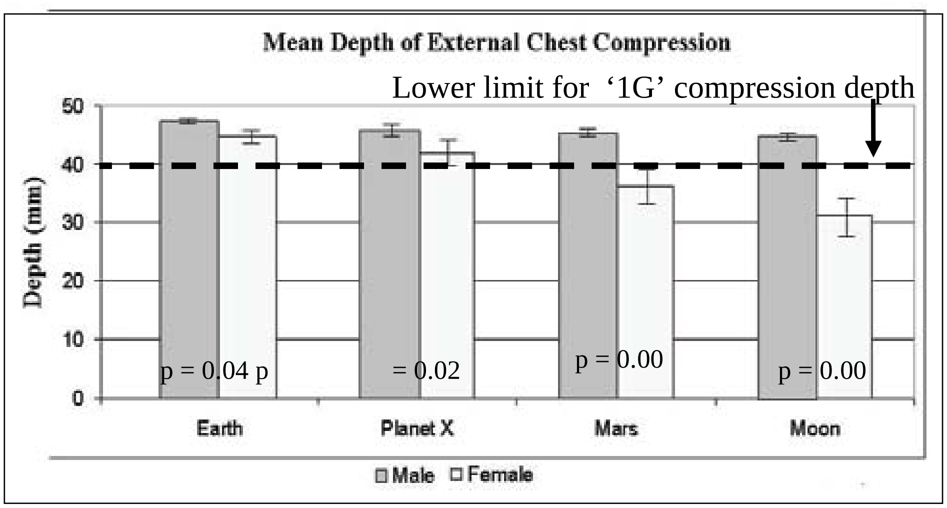


Figure 6. Chest compression depth in four gravitational environments. Values along Y axis represent chest compression depth in millimetres.



satisfactory compressions compared to males (p=0.03). Additionally, out of the seven subjects who were unable to complete the three minutes of CPR due to exhaustion, five were female. The authors also observed a good correlation between the number of correct chest compressions performed over the study period and subject height (first period: Spearman r=0.40, p=0.01; second period: r=0.32, p=0.05) and weight (r=0.51, p=0.0008; r=0.48, p=0.002). They did not find a significant correlation between CPR performance and age.

In another study of Larsen, et al. (2002), 20 medical students were observed while carrying out four cycles of 15 external chest compressions followed by two ventilations on a manikin. The data illustrated that the depth of compression, time to peak compression, and the regression slope of compression vs. compression number were significantly related to height and weight of the rescuer. Therefore, as well as ascertaining the role of gravity, the effects of these biometric factors (mass, height, gender, and physical fitness) on performance of

Table 3. p Values for male results

Male	p values					
	1G x Px	1G x Ma	1G x Mo	Px x Ma	Ma x Mo	Px x Mo
Mean DCC (mm)	0.119	0.101	0.087	0.274	0.089	0.103
Mean FCC (crp/min)	0.136	0.189	0.196	0.248	0.461	0.187
Elbow Flexion(º)	0.178	0.001	0.000	0.000	0.032	0.000
Load Variation (Kg)	-	-	-	0.003	0.063	0.000
HR Variation (bpm)	0.018	0.002	0.009	0.079	0.449	0.195

DCC: Depth of chest compression; FCC: Frequency of chest compression; HR: heart rate; Px: Planet X; Ma: Mars; Mo: Moon

Table 4. p Values for female results

Female	p values					
	1G x Px	1G x Ma	1G x Mo	Px x Ma	Ma x Mo	Px x Mo
Mean DCC (mm)	0.095	0.018	0.000	0.016	0.015	0.001
Mean FCC (crp/min)	0.160	0.053	0.089	0.076	0.243	0.049
Elbow Flexion (º)	0.063	0.005	0.002	0.004	0.297	0.005
Load Variation (Kg)	-	-	-	0.000	0.051	0.000
HR Variation (bpm)	0.000	0.000	0.000	0.420	0.135	0.177

DCC: Depth of chest compression; FCC: Frequency of chest compression; HR: heart rate; Px: planet X; Ma: Mars; Mo: Moon

chest compression were also ascertained and quantified during the study.

This current study looked at evidence for the effectiveness of the terrestrial method in producing adequate compression depths and rates in these low-gravity environments, and also looked at how these parameters vary with gravity and body mass of the caregiver to allow the determination of “critical gravity” and “critical body mass” where compression depths become effective.

The body suspension device used successfully reduced subject body weight to values close to that expected to be encountered in hypogravity environments such as those of the Moon and Mars. Subjects were able to monitor their depth and rate of chest compressions effectively, a point highlighted by the fact that mean rates of compression were maintained during all conditions with only one exception.

The results clearly show that appropriate depths and rates of chest compression do appear to be possible for stronger/heavier (in this case, male) rescuers when subjected to hypogravity in the region of 0.17Gz to 0.7Gz. Lightweight subjects (predominantly female), however, were unable to achieve adequate chest compression depth in the Lunar and Martian hypogravity environments.

The present study suggests that effort required to perform adequate external chest compressions is greater, as demonstrated by the increase in heart rate following three minutes of CPR. Furthermore, the angle of elbow flexion increases with reductions in hypogravity, suggesting that chest compression force increasingly comes from the arm/shoulder musculature as body weight is reduced. Upper arm strength may therefore be important in achieving adequate external chest compressions in hypogravity.

A study conducted by Ganeshanantham, Russomano, Dalmarco, Calder, Evetts, and Cardoso (2006) aimed to illustrate the role of elbow extension in the maintenance of an effective ECC in a lunar environment. Eight healthy male subjects performed ECC while in the same body suspension device. After training, subjects followed a protocol of 30 ECC to 6s of rest (the time needed for the performance of two ventilations) (AHA, 2005) for 3 minutes. ECC adequacy was assessed by rate and depth of compressions. During compressions, subjects were instructed to maintain the straight arms terrestrial method of performing ECC, with arm braces employed to impair the subject's ability to extend and flex the elbows. Results showed a lower mean elbow angle variation of 5.95 ± 3.4^{o} ($p<0.01$) and concomitantly lower mean depth of compression that fell below the clinical threshold (37±2.9cm, $p<0.01$) in comparison to the present study findings. It can thus be concluded that elbow extension and muscular work of the upper limbs play a vital role for ECC performed in a Lunar environment, without which the current CPR standards cannot be maintained.

CONCLUSION

The PUCRS Microgravity Laboratory body suspension system can reproduce hypogravity conditions such as those expected to be experienced on Mars and the Moon. The results of this study of terrestrial CPR in hypogravity suggest that these conditions lead to an alteration of the manner in which the terrestrial method of external chest compression is performed. It appears that the rescuer arm and shoulder muscular effort by means of greater elbow flexion and extension is increased to counter the decrease in body weight. The study also indicates, however, that smaller, weaker rescuers (in this case, the female subject group) may not be able to apply sufficient elbow extension force to achieve adequate chest compression under these conditions.

REFERENCES

American Heart Association. (2000). Guidelines 2000 for cardiopulmonary resuscitation and emergency cardiac care.

Ashton, A., McCluskey, A., Gwinnutt, C.L., & Keenan, A.M. (2002). Effect of rescuer fatigue on performance of continuous external chest compressions over 3 min. *Resuscitation, 55*(2), 151–155.

Cummins, R.O., Eisenberg, M.S., Hallstrom, A.P., & Litwin, P.E. (1985). Survival of out-of-hospital cardiac arrest with early initiation of cardiopulmonary resuscitation. *Am J Emerg Med, 3*(2), 114–119.

European Space Flight History (ESA). (2006). Retrieved from http://www.esa.int

Evetts, S. (2004). *Non-terrestrial basic life support*. Proceedings of the Conference on abs Inaugural UK Space Medicine Day. Leicester, UK.

Evetts, S., Evetts, L., Russomano, T., Castro, J.C. & Ernsting, J. (2003). A validity assessment of a new method of performing single person CPR in microgravity. Medicine and Mobility Space Biomedicine Congress, Berlin.

Evetts, S., Evetts, L., Russomano, T., Castro, J., & Ernsting, J. (2004). *An assessment of the feasibility and effectiveness of performing CPR during microgravity*. Proceedings of the Annals of the 75th AsMA Annual Scientific Meeting: Frontiers in Aerospace Medicine, Anchorage, Alaska.

Evetts, S., Evetts, L., Russomano, T., Castro, J.C. & Ernsting, J. (2005). Basic life support in microgravity: Evaluation of a novel method during parabolic flight. *Aviat Space Environ Med, 76*, 506–510.

Exploration of Mars and the Moon. (2006). Retrieved from http://spaceflight.nasa.gov/mars

Future of Human Space Flight. (2006). Retrieved from http://www.space.com/spacetourism

Ganeshanantham, G., Russomano, T., Dalmarco, G., Calder, A., Evetts, S., & Cardoso, R.B. (2006). Elbow angulation—Key to performing effective lunar CPR. *Proceedings of the 3rd UK Space Medicine Conference abs*, Leicester, UK

Hamilton, D. (2003). Wyle/NASA flight surgeon. NASA Lyndon B. Johnson Space Centre. Personal Communication.

Handley, A.J., Monsieurs, K.G., & Bossaert, L.L. (2001). European Resuscitation Council 2000 guidelines for adult single rescuer basic life support. *Resuscitation, 48*, 199–205.

Japan Space Flight History. (2005). Retrieved from http://www.space.com/missionlaunches/ap_050406_japan_moon.html

Jay, G.D., Lee, P., Goldsmith, H., Battat, J., Maurrrer, J., & Suner, S. (2003). CPR effectiveness in microgravity: Comparison of three positions and a mechanical device. *Aviat Space Environ Med, 74*(11), 1183–1189.

Johnston, S. (1998). NASA flight surgeon. NASA Lyndon B. Johnson Space Center. Personal communication.

Johnston, S.L., Campbell, M.R., Billica, R.D., & Gilmore, S.M. (2004). Cardiopulmonary resuscitation in microgravity: Efficacy in the swine during parabolic flight. *Aviat Space Environ Med, 75*(6), 546–550.

Larsen, P.D., Perrin, K., & Galletly, D.C. (2002). Patterns of external chest compression. *Resuscitation, 53*(3), 281–287.

Lucia, A., et al.. (1999). The importance of physical fitness in the performance of adequate cardiopulmonary resuscitation. *Chest, 115*(1), 158–164.

Paradis, N.A., et al. (1989). Simultaneous aortic, jugular bulb and right atrial pressures during cardiopulmonary resuscitation in humans: Insight into mechanisms. *Circulation, 80*, 361–368.

Ritter, G., et al. (1985). The effect of bystander CPR on survival of out-of-hospital cardiac arrest victims. *Am Heart J, 110*(5), 932–937.

Safar, P. (2004). Closed-chest cardiac massage. *Acta Anaesthesiol Scand*, Suppl 9, 125-128.

Sarkar, S. (2004). Cardiopulmonary resuscitation in Lunar and Martian gravity fields. Aviation, Space and Environmental Medicine, 75(4), B40.

Telemedical Emergency Management on Board the International Space Station: Final Report. (2004). Retrieved from http://www.e-ms.de/de/reports/temos2004/TEMOS.pdf

White House Press Secretary. (2004). President Bush announces new vision for space exploration program. Retrieved from http://www.whitehouse.gov/news/releases/2004/01/20040114-1.html

Wiggington, J.G., Miller, A.H., Benitez, F.L., & Pepe, P.E. (2005). Mechanical devices for cardiopulmonary resuscitation. *Current Opinion in Critical Care, 11*(3), 219–223.

KEY TERMS

Cardiac Arrest: Also known as cardiorespiratory arrest, cardiopulmonary arrest, or circulatory arrest, cardiac arrest is the abrupt cessation of normal circulation of the blood due to failure of the heart to contract effectively during systole.

Cardiomyopathy: Disease of the heart in which the heart muscle becomes inflamed and doesn't work as well as it should.

CPR: Cardiopulmonary resuscitation.

Haemodynamic: The dynamic regulation of the blood flow in the brain.

Morbidity: Death.

An Exploration of Demographic Inconsistencies in Healthcare Information Environments

Larbi Esmahi
Athabasca University, Canada

James W. Edwards
Agfa Healthcare, Canada

Elarbi Badidi
United Arab Emirates University, UAE

INTRODUCTION

Healthcare providers typically use technology in a ubiquitous manner. They choose to rely upon the task-specific capabilities of a specific system, rather than the integration of all systems in the solution space. A typical example of this would be the advancement of ultrasound technologies for 3-D modeling. Whilst the capabilities of the imaging solutions are highly beneficial to the individual's unique needs for healthcare, the integration of this system into a greater solution, or system-of-systems, is often overlooked (Maier, 1998; Maier, 2005). As such, the demographic data accompanying the information may be in a format inconsistent with the requirements for the electronic medical record. Workflow Management Systems attempt to simplify and control the primary data entry methods to the data environment (Choennia, Bakkera, & Baetsa, 2003; Graeber, 1997). However, the design of these systems focuses on an interpretation layer to gather data from varying input sources, including voice and text. There is no guarantee that this information will propagate throughout the system-of-systems, as required. Common information protocols such as Digital Imaging and Communications in Medicine (DICOM) (NEMA, 2004) and Health Level Seven (HL7.org, 2006) provide a framework by which medical data is communicated. However, these standard protocols are subject to interpretation that allows for a high degree of variance in the presentation of data. This creates problems of data redundancy, and confuses the authoritative provisioning of data with referential copies.

The electronic healthcare record (EHR) is a merged presentation of the information obtained through various systems in the healthcare enterprise (Hasselbring, Peterson, Smits, & Spanjers, 2000). Imaging and demographic data contribute to the contents. Additionally, the physician comments, markup, and modality reports also contribute to the record. These individual pieces of the EHR are drawn from many disparate systems throughout the enterprise. The record information system (RIS) or alternatively a health information system (HIS) will be the primary demographic repository for patient information. As such, this text and contextual information will be used as the primary data source for downstream comparison. The Picture Archive and Communication System (PACS) will be used as a repository for imaging data (Miltchenko, Pianykh, & D'Antonio, 2002).

As the global economy and global business practices increase, individuals tend to be more transient. Their medical histories are an important precursor to successful health care provisioning. Contingent upon this medical history record is the successful integration of data from varied sources, including the mobile patient's information. This information must be portable, presentable, and independent of the initial data program from which it was obtained. By far, the greatest concern relates to data inconsistency and subsequent inaccuracy in an environment of disparate systems.

In the next section of this article, we will discuss the Health Level Seven communication protocol. In the third section, we will discuss the protocol DICOM. The application of these protocols into solutions for healthcare information repositories is presented in the fourth section, where both PACS and RIS systems are discussed. The problem of Data Disparity that causes inconsistencies in the intersystem communications will be discussed in the fifth section. The design of the XML Bus solution will be presented in the sixth section. The successes and shortcomings of this design are discussed in the conclusion.

HEALTH LEVEL SEVEN

The HL7 protocol provides an information standard for communicating patients and sites level information into and out of a health information system within the system-of-systems model (HL7 Canada, 2002; HL7.org, 2006; Neotool, 2007). HL7 allows for a loosely framed communication standard between medical IT systems. Without the provisioning of this ANSI standard, interoperability between systems was performed in an ad hoc manner.

The HL7 standard is based on the concept of events referred to as triggers (HL7, 2006, 2007). Each of these triggers fires a process that effects some particular communication of information. For example, the relocation of a patient to a new hospital would require the dissemination of information concerning that location throughout the healthcare system-of-systems. This event would be one of such triggers. A trigger may relate to a single record, as in this example, or it may relate to multiple records. A patient may have a name change due to marriage. This would require an update to all existing records in the information system to ensure data consistency.

Each trigger is composed of segments. The segment is a variable length field separated by a discernable character, such as a comma or pipe. In a manner similar to comma separated values (CSV) files, data can be delineated and also referenced externally though API functions. The content of each segment shall contain ASCII characters. The HL7 committee defines the acceptable character set as "The ASCII displayable character set (hexadecimal values between 20 and 7E, inclusive) is the default character set unless modified in the MSH header segment" (HL7, 2006). The HL7 trigger is represented in Figure 1.

Each field may contain multiple subfields, delimited by a caret character. In the case of a name, |*DOE^JOHN^^^^*| the content between the two pipe symbols is the field of data, whereas the data *DOE* and *JOHN* are considered subfields. The MSH is the message header segment, providing basic instructions on the contents within the trigger message.

This example also makes reference to three other key items: the *PID* that is the patient identifier, the EVN identifying that an event is included in this message, and the PV1 filed, which is followed by information about the patient's stay at the hospital.

One of the important contents within this message is the event type. As per the example, the message type provided is A04. HL7 defines many different message types, each with unique content. There are 51 defined Admission, Discharge or Transfer (ADT) events for HL7 in the V2.5 standard. The following table describes some of these events.

Without elaborating further on some of the many messages available to the HL7, it becomes clear that many distinct attributes are communicated in this protocol. Static entities such as Patient ID, Accession number, and event dates are provided in the same framework as dynamic information, such as location. The messaging protocol is also intended to communicate changes in state, through the provisioning of state-in-time attributes that have an expectation of permanence (Evola, 1997). These attributes, such as Patient ID, may have referential integrity for a site perspective. However, in the conglomerate or regional model of healthcare, that identifier now becomes nonreferential from a global viewpoint, and as such, introduces disparity at a higher system-of-systems level.

Figure 1. Sample HL7 trigger (Interfaceware, 2006)

```
MSH|^~\&|EPIC|EPICADT|SMS|SMSADT|199912271408|CHARRIS|ADT^A04|1817457|D|2.3|
EVN|A04|199912271408|||CHARRIS
PID||0493575^^^2^ID 1|454721||DOE^JOHN^^^^|DOE^JOHN^^^^|19480203|M||B|254
E238ST^^EUCLID^OH^44123^USA||(216)731-4359|||M|NON|400003403~1129086|999-|
NK1||CONROY^MARI^^^^|SPO||(216)731-4359||EC|||||||||||||||||||||||||
PV1||O|168   ~219~C~PMA^^^^^^^^^^||||277^ALLEN FADZL^BONNIE^^^^|||||||||
||2688684|||||||||||||||||||||||||199912271408||||||002376853
```

Table 1. HL7 message types

Message Type	Definition
Admit/visit notification (event A01)	Assign a bed to an admitted patient
Transfer a patient (event A02)	Patient changes their physical location
Discharge/end visit (event A03)	Patient leaves the facility
Register a patient (event A04)	Patient is brought into the facility, but not given a bed
Pre-admit a patient (event A05)	Patient is not admitted (In the facility), but demographic information is registered in the information system prior to their arrival for procedure.
Change an outpatient to an inpatient (event A06)	If a patients status changes, normally due to severity, they will be given a new location at the facility
Change an inpatient to an outpatient (event A07)	As per A06, defines a state change in the patient location.
Update patient information (event A08)	Represents a change to any of the demographic information related to a patient. This may be name, address, or other field.
Patient departing tracking (event A09)	Patient in transit outbound
Patient arriving - tracking (event A10)	Patient in transit inbound

DICOM

DICOM exists as an information distribution mechanism for medical devices (Ratib, Ligier, Rosset, Staub, Logean, & Girard, 2000). It is important to note that DICOM can be viewed as a technical protocol for layer seven communications or a solution to information exchange.

The DICOM protocol was developed in a joint effort between the American College of Radiology (ACR) and the National Electrical Manufacturers Association (NEMA, 2004). The protocol was introduced as a hardware interface, a minimum set of software commands, and a consistent set of data formats for communication between medical devices. This standard was intended to provide a framework to an emerging technology, medical information technology, allowing the sharing of data amongst Picture Archive Communication Systems and compliant sources.

The DICOM protocol is used by a multitude of devices within the hospital enterprise for information transfer, sharing, and coordination (Jung, 2005). Some of the systems that rely on the protocol include modalities (Ultrasound, X-ray, CR, CT, and so on), Picture Archive Communication Systems, Film Printers, and Digitizers (media conversion systems) (Iacucci, Nielsen, & Berge, 2002). Although the implementation of the protocol is proprietary for each manufacturer, the standards as provided by the NEMA organization assist in ensuring general interoperability and definition of expectation for each DICOM function.

The first layer of protocol discussion relates to the negotiation of capabilities between the acquisition device and the subsequent receiver. Two concise roles are defined within DICOM to elucidate the responsibilities of each participant in the information exchange: the Service Class User (SCU) and the Service Class Provider (SCP). The SCU is the requestor of services for the purposes of information storage. In essence, it has information, and solicits a predefined receiver to accept this information. The SCU negotiates the acceptance of the sender's data, typically in one or more ways (Revet, 1997).

The storage transmission starts with a request from the SCU. The SCU identifies itself and asks for a specific function, through a composite request. The request consists of an object and an associated action. For example, a C-STORE message may consist of a request for data storage, and a subsequent object, such as an ultrasound image. The SCP, or storage provider, accepts the association request from the resource, assuming the storage host trusts it.

As illustrated in Figure 2, the bidirectional request-response behavior is typical of the DICOM protocol. The SCP not only accepts the initial request, but also responds with a list of capabilities for that request. The response consists of a list of Service Object-Pair (SOP) classes, reflective of the storage ability of the SCP system. Also included in this request-response exchange is the choice of transfer syntax used for communication. Since there is a multiplicity of available syntaxes in the DICOM standard, some systems require the presence of an explicit value representation in the message. Others rely on an implicit understanding between data dictionary implementations, as provided partly by the DICOM standard, and partly by the SCP developer (McCormick, 2006).

DICOM data is broken into six major groups: File, Study, Acquisition, Image, Object, and Presentation. Each of these groups is well defined in terms of permissible content by the DICOM committee, as published in the DICOM standards and Supplements. Seven major groups are used to categorize DICOM Data elements. Group 0002 identifies file level attributes. Group 0008 identifies attributes relating to the study. Group 0010 identifies the patient attributes. Group 0018 relates to the acquisition of specific attributes related to the image. Group 0020 contains object level information, including the hierarchal relationship attributes that allow the linking of this image or object to the other objects in the series, study, and patient. Group 0028 elements provide display information crucial to the proper visual presentation of the image. Medical software uses this information to apply default settings. Group of Private Data Elements (odd numbered groups), allow the vendor to include information that can be used in a proprietary manner (i.e., to display the images, in case that vendor's system is part of the total System-of-Systems solution).

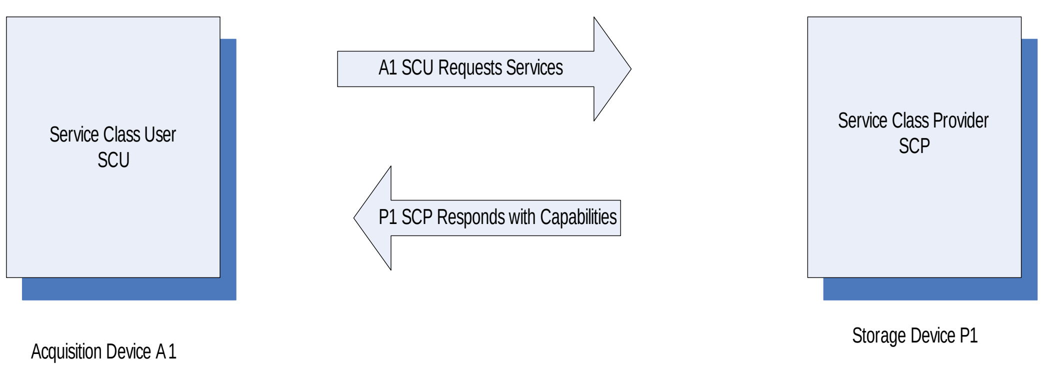
In summary, the DICOM protocol allows not only for the communication of patient, study, series, and image level information, it also defines the mechanism by which that information may be exchanged (Kolesnikov, 1997; Kolesnikov, Kelle, & Kauranne, 1997). This data is the primary source for study information as it related to the patient.

PACS AND RIS

PACS systems are built as a central repository for DICOM information (Gell, 2000). They are typically deployed as a means of replacing typical film based media. Through the implementation of the Storage Class User (SCU) and Storage Class Provider (SCP) mechanisms defined in DICOM, the PACS system receives images for later retrieval. PACS provides an electronic means of long-term storage. This storage serves as the integrated lifecycle management mechanism for a continued inbound stream of data (Miltchenko et al., 2002). The PACS is also the recipient of messages from the RIS. These messages represent unique events, such as a patient name change. Upon reception of such events, the PACS system is expected to start a process that will eventually update the demographic data using the previous name to the new patient name. The PACS answer to subsequent queries related to the updated patient records is subject to the proper interpretation of the data, as identified through the use of explicit data types, as well as the data dictionary. In fact, the use of DICOM data dictionaries allow for explicit metadata definition of data streams received. In the case of PACS, data dictionaries assist in defining the understanding of data content, both by type and content (Kennedy, Seibert, & Hughes, 2000).

Health Information Systems (HIS) organize demographic records for the purpose of centrally collecting a repository of electronic information. Radiology Information Systems (RIS) build upon this purpose for the radiology patient by scheduling orders. As information repositories, these systems (HIS and RIS) are constantly updated with both new and changing information. RIS systems can also provide worklists to modali-

Figure 2. SCU-SCP communications



ties containing patient names and other demographic information of importance. In order for modalities to communicate outwards to information systems, they rely upon the DICOM protocol (Miltchenko et al., 2002; Tornhola, 2004). This may require the HIS system to accept DICOM events. Alternatively, data translation systems occasionally transform DICOM messages to the appropriate HL7 data construct, normally understood by the HIS system. The DICOM standard assists PACS implementations in providing a referential target, but only provides DICOM consistency of the data that is independent of the RIS. The RIS always relies on the HL7 protocol and its messaging mechanism for informing the PACS of changes, as well as receiving events from DICOM devices.

The messaging mechanism between different systems offers the translation medium that serves to both interpret and re-represent the data. However, the interpretation is based on an understanding of the message context inbound. This is a highly subjective interpretation. Resultantly, events and interpretation can result in inconsistency in this messaging based approach.

CAUSES FOR DATA DISPARITY IN THE SYSTEM-OF-SYSTEMS APPROACH

As discussed in the previous sections, the event and messaging framework used in the health enterprise system does not warrant the consistency between different instances or parts of the patient records. In this section, we will examine some of the events that may end with data disparity in the health enterprise system.

Data Silos in Healthcare Information Systems

Protocols such as DICOM and HL7 exist for the purposes of exchanging information within the PACS and RIS information silos in the hospital enterprise. These protocols ensure data confidence for downstream systems, but are not designed to provide referential data cross system in the system-of-systems model. These protocols provide a framework by which medical data is communicated, but are subject to strict programmatic interpretation. System implementations interpret the protocol standards in unique ways, allowing for a high degree of variance in the presentation of data from multiple independent data sources. This creates problems of data redundancy, inaccuracy, and confuses the authoritative provisioning of data with referential copies. By far the greatest concern relates to data inconsistency and subsequent inaccuracy in an environment of disparate systems. Unique data keys used for data alignment or referential integrity may be drawn from human assumptions that are faulty.

Inconsistent Record Identifier for the Patient's Data

Name change: In the system-of-systems, the patient name is always used as a key or part of the key for querying the patient data. Patients sometime change their names, especially after a marriage or divorce. When such an event occurs, demographic references strongly tied to that person's medical history will need to be changed or updated in order to insure the continuous access to relevant historical data.

In order to synchronize a patient's name throughout the system-of-systems enterprise, an HL7 N-event entitled "Patient Update" needs to occur. This type of event will ensure that all records containing "patient name" are updated to reference "the new patient name."

Legacy system for scanning film and paper data: Many of the enterprise records for patients in the hospital exist on documents. Admission, discharge, and transfer records may exist solely on nonelectronic medium. These information sources must be manually updated each time an event occurs, such as a patient name change. More information exists in the way of medical film. Not only is the image data present, but often the demographic information for the patient is burnt into the film.

OCR technologies can be helpful in dealing with the shear volume of ADT records for the patients. However, the electronically introduced data will need to be reconciled with the existing HER records.

Data Inconsistency Generated by Automatic Triggers or Cascading Events

Study cancel: On a less frequent basis, patient examinations are not conducted. This could be due to a myriad of causes, including patient absenteeism, acquisition device failure, as well as general infrastructure failure. In any of these events, a decision is made to cancel the acquisition of data relevant to the patient. In these cancellation scenarios, an N-event message is then scheduled. However, when the message is received by the PACS system, the study or appropriate identifier may not be present, as defined in the N-Event. This may be caused if the study acquisition has not yet begun. Also, the billing system may not receive the correct event-message. As such, we have an event-ordering problem.

Study split or merges: Data flow allows for the division and recombination in the process of medical diagnosis. A patient may have had a very long CT exam that extends into body parts outside the area of expertise for the radiologist. As such, it may be beneficial, both medically and financially, to split the study into two distinct studies for the patient. The inverse may also occur, resulting in a study merge. Study or series acquisition may be interrupted for one or more reasons, including equipment failure, resource contention, or multiple system acquisition. In any of these events, two or more distinct studies may need to be merged into a single study record.

A question arises regarding the synchronization of data between these models. It is unclear whether one or more IS records in this environment would exist for each study. As such, we encountered a many-to-one disagreement. Billing systems would need to be correctly informed that only one, albeit larger, procedure was performed. Also, only one read or transcription should have occurred in this model.

Data Exchange and Transfer of Patient Records

In some medical situations, a patient will need to be relocated from one hospital environment to another. In this dataflow scenario, the patient's records may be imported to the new environment. Alternatively, a new record may be created at the newest site. Also, the imaging data is sometime transmitted to the DICOM SCP of the receiving site's PACS. The transmitted and imported data needs then to be reconciled against the new site's information system in order to correlate records to demographic data, so that any patient movement or other procedures can be referenced to the same patient.

Character Set Translation and Legacy Systems Tagging

ISO_IR disagreement and database character translation: ISO character sets refer to the specific procedure for mapping a sequence of bytes to a sequence of characters. The DICOM standard includes a specific tag necessary to identify the character set used in the header. This is of extreme importance to the demographic consistency of the data interpretation and action throughout the enterprise.

By way of example, consider a query for all patients whose last name starts with Ł (from ISO 8859-2, decimal 163). If the system were not capable of understanding this character, perhaps through a limited understanding of only ISO_8859-1, then this character would escape the query, possibly with undesirable results. Worst yet, any patient name update including the character Ł would also be problematic.

Private DICOM tagging: Many modality vendors choose to include proprietary information in the DICOM header of the images and objects they transmit. This information is only of use to a system that is programmed to take advantage of it. It allows for a tighter sub-system-of-system model, but at the expense of the greater solution.

Time Synchronization

Time is a critical component to a successful diagnosis. All members of the health enterprise solution normally have their own time keeping facility. However, in the system-of-systems, the use of time stamps is critical to insure the coordination of transactions and order of precedence. These time stamps should be synchronized within the whole system-of-systems.

Other Technical Issues

Data reintroduction for disaster recovery purposes: In the event of system failure, one or more entities within the system-of-systems will need to be repopulated with original data. Upon completion of this, a synchronization effort will need to occur between it and one of more members within the environment. This activity requires coordination of the various information systems within the hospital enterprise.

E

RIS feed interruption: The RIS messaging mechanism is a necessary prerequisite in the system-of-systems for data continuity. Upon failure of any RIS system in this environment, the update events will not properly register for all receiving systems in the enterprise, resulting in an environmental data inconsistency.

AN XML BUS FOR RESOLVING SOME OF THE DATA DISPARITY ISSUES

One of the prevalent issues in message-oriented systems is defining one or more protocols by which messages can be processed. Adherence to standards, bridging between protocols and configurations all raise issues in a multiple systems environment. In the healthcare enterprise, the interdependency of systems for successful patient care only serves to exasperate the situation. XML and semantic Web technology presents a great potential for solving some of the disparity issues (Heitmann, 2002; Pianykh, Castañeda-Zuniga, W. D., D'Antonio, & Tyler, 2001a; Pianykh, Miltchenko, & Castaneda-Zuniga, 2001b; Tornhola, 2004). In this section, we present a solution based on the use of an XML bus for data transfer. Figure 3 presents the XML bus design from a dataflow perspective.

The data bus would allow for a common data conduit. The connection to this conduit would be accomplished through adapters that provide common data typing. Each connection into the data bus would require an adapter that allows for consistent conversion from source protocol to XML.

DICOM to XML Conversion

The adapter, as shown in the arrow in Figure 3, is a translation engine that understands both XML and the healthcare-specific protocol, in this case, DICOM. Using open source tools such as DCM4CHE, conversion of DICOM to XML is readily possible.

The following code is a truncated output of the dcm2xml.jar transformation performed on a sample DICOM file.

HL7 to XML Conversion

Version 3.0 of the HL7 standard is a pure XML implementation (Heitmann, 2002). As such, the major concern relates not to the transformation of the trigger, but to the interpretation. This is somewhat addressed in the Reference Information Model of HL7 3.0. In versions prior to 3.0, a methodology is necessary to interpret the trigger events as XML. The HL7 organization has devised rules in order to properly encode the HL7 triggers in XML. These rules would enable an adapter framework to implement commonality of meaning across the data bus, as shown in Figure 3.

Modality to XML Data Bus Adapter

There is commonality between the DICOM to XML Adapter, and the Modality to XML adapter. Both use DICOM as a protocol for communication, and as such, may rely partially on the implementation details for each other. Modalities also have unique features, such as Modality Performed Procedure Steps (MPPS), which must also be taken care of in any successful implementation

Figure 3. XML data bus

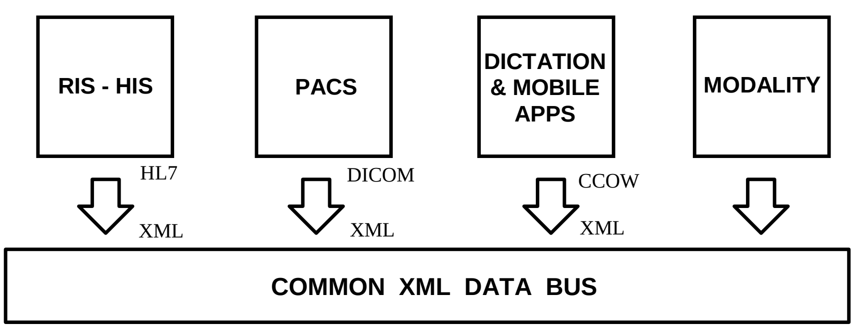


Figure 4. Excerpt from DCM2XML.jar output


```
<?xml version="1.0" encoding="UTF-8"?>
<dicomfile>
<filemetainfo>
<attr tag="00020000" vr="UL" pos="132" name="Group Length" vm="1" len="4">196</attr>
<attr tag="00020001" vr="OB" pos="144" name="File Meta Information Version" vm="1" len="2">1\0</attr>
<attr tag="00020013" vr="SH" pos="290" name="Implementation Version Name" vm="1" len="16">eFilm/efDICOMLib</attr>
<attr tag="00020016" vr="AE" pos="314" name="Source Application Entity Title" vm="1" len="18">            &#19;</attr>
</filemetainfo>
<dataset>
<attr tag="00080005" vr="CS" pos="340" name="Specific Character Set" vm="1" len="10">ISO_IR 100</attr>
<attr tag="00080008" vr="CS" pos="358" name="Image Type" vm="2" len="16">DERIVED\PRIMARY </attr>
<attr tag="00080012" vr="DA" pos="382" name="Instance Creation Date" vm="1" len="8">20010109</attr>
<attr tag="00080013" vr="TM" pos="398" name="Instance Creation Time" vm="1" len="6">095618</attr>
<attr tag="00080016" vr="UI" pos="412" name="SOP Class UID" vm="1" len="26">1.2.840.10008.5.1.4.1.1.1</attr>
<attr tag="00080018" vr="UI" pos="446" name="SOP Instance UID" vm="1" len="40">1.3.51.5145.5142.20010109.1105627.1
.0.1</attr>
```


(Noumeir, 2005). The modality adapter must be able to communicate these specific events directly to the data bus. Any subsequent queries upon the data bus would then be able to accurately provide procedure start and procedure end dates and times through the Modality to XML adapter shown in Figure 3.

Context to XML

Many applications rely on the Clinical Context Object Workgroup (CCOW) for interoperability notification. This standard and an extension of the HL7 framework allows singular CCOW-aware applications to interact with other CCOW-aware applications for the purposes of information sharing (Grimson, Stephens, Jung, Grimson, Berry, & Pardon, 2001). This standard is extremely beneficial in the system-of-systems approach, as it represents a common notification mechanism for changes in status. It also has value in voice dictation systems as a means of coercing dictation information into the patient framework through the identification of common attributes, such as patient name and study instance-unique identifiers.

Proof of Concept

The following interface provides a rudimentary proof of concept for the Enterprise Healthcare Record Viewer application. The intent of this application is not to commit a full implementation, but rather to provide a framework by which further research may be committed.

Figure 5. Enterprise healthcare record viewer prototype

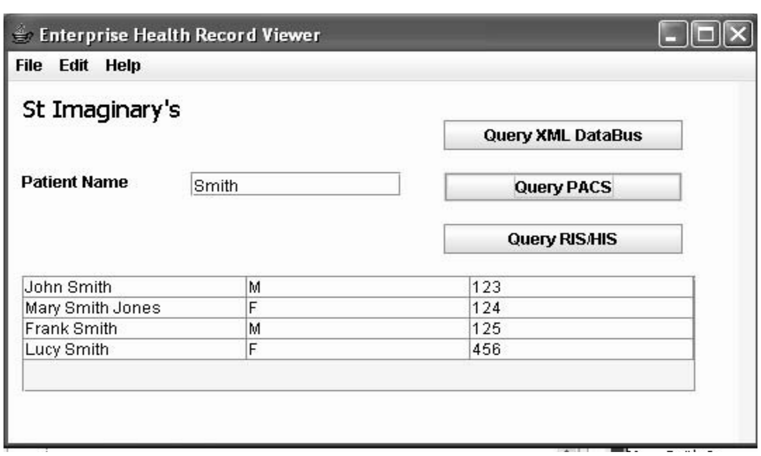


Figure 6. DICOM response from EHR application C_Find

Grp	Elmt	Length	Description	VR	VM	Value
0008	0052	8	query_retrieve_level	CS	1	"PATIENT"
0010	0010	6	patient_name	PN	1	"Smith"
0010	0020	2	patient_id	LO	1	"*"
0010	0040	2	patient_sex	CS	1	"*"

The EHR interface provides three query mechanisms. The XML Data Bus function performs a direct SQL query to the data repository, as stored through a future XML to SQL transformation. The second interface queries the PACS SCP through a standard C_FIND. Figure 6, an excerpt of the DICOM conversation ensuing from that query, details the Group and Element data transferred.

The last query mechanism queries the RIS system. Connectivity to the RIS is accomplished through an HL7 Query message.

CONCLUSION

Medical information systems exist as information silos within the global health enterprise, making the task of designing an effective and accurate Enterprise Health Record Viewer very difficult. Each healthcare enterprise system in the system-of-systems acts as a referential entity for the information it contains, but only for the specific domain for which it is specialized. For each system in the enterprise, downstream distributed applications typically rely on the one specific data silo as a means of identifying and enacting decisions within an application framework. Whilst this ensures one to one consistency of data between the application and a centralized data repository, it does not ensure consistency within the enterprise.

In the enterprise health record, two systems were found to be of importance. Radiological clients communicated primarily with a PACS system via DICOM as a means of obtaining study level data for patients. Radiology workstations communicated with RIS systems via HL7 in order to record patient-level attributes. Data consistency was often accomplished through the implementation of translation systems converting one protocol to another. In fact, protocols such as DICOM, CCOW, and HL7 exist as standards for information sharing, but the exact implementation of message conversion from one standard to another is often an area of information redundancy. Data conversion solutions were also viewed as a possible cause for information loss or inaccuracy due to the incongruity in frameworks between the systems and their protocols. In the translation and representation between systems in the enterprise, data conversion is not always possible. Enterprise inconsistencies also exist as a direct result of data workflow. HIMMS and IHE are leading the way in standardization of data consistency and authority in the healthcare enterprise by defining information pathways and messaging content within the medical information environment.

The development of an Enterprise Record Viewer illuminated the necessity for authority of data. Although protocol translation was seen as a contributing factor to inconsistencies, the production of a solution became reoriented around the creation of a central authority in order to ensure consistency. This hierarchal approach avoided the concern over data consistency in the enterprise by creating another independent silo for data display. Connectivity to the RIS and PACS systems allowed the viewer to query independently of the XML Data Bus in order to allow alternative means of exploring data that may not have been messaged properly.

The production of the framework for the XML data bus is representative of a methodology or approach for an Electronic Healthcare Record. Further effort is necessary to expand upon this singular solution to resolve the data disparity issues driven by the system-of-systems approach in the hospital enterprise. The XML data bus design provides a framework for solving some of the identified causes for data disparity. Patient name changes are addressed through the transmission of this data to the bus. All queries can then be programmed to query via the bus. Similarly, patient moves and study cancellation messages can be communicated into the bus for dissemination and subsequent query. While the bus design alleviates these areas by creating a singular entry point for enterprise systems, other causes for data inconsistency cannot be solely addressed through the XML data bus design. Consistencies in data dictionaries, ISO_IR disagreement and private DICOM tagging practices require future research. These areas may benefit from an assessment of the W3C semantic Web design. Another concern remains with data movement and event ordering. Issues such as time synchronization and RIS message repetition may benefit from an asynchronous transactional approach as implemented through the use of a message queuing. Legacy film based scanning practices would benefit from an examination of form based workflow technologies, and would integrate more readily into an XML construct such as the proposed XML data bus.

Future implementations are recommended to focus on the remote user's requirements as a means of both centralizing data, and also on providing a readily accessible, consistent, and lightweight viewpoint to data. This latter objective will facilitate a successful adoption through the consistent presentation of patient demographic data. As Scotland, England, Estonia, Canada and others begin the process of providing centralized socialized healthcare through the amalgamation and adoption of countrywide healthcare solutions, Enterprise Healthcare Record Viewers and identification of data authorities inside the system-of-systems healthcare enterprise will become critical to the successful provisioning of healthcare to patients.

REFERENCES

Choennia, S., Bakkera, R., & Baetsa, W. (2003). On the evaluation of workflow systems in business processes. *The Electronic Journal of Information Systems Evaluation, 6*(2), Paper 5. Retrieved February 12, 2008, from http://www.ejise.com/volume6-issue2/issue2-art5.htm

Evola, R. L. (1997). *Patient administration*. Article 3, in Health Level Seven, Version 2.3. Retrieved February 12, 2008, from http://www.dmi.columbia.edu/resources/hl7doc/hl72.3/ch3.pdf

Gell, G. (2000, February 15). PACS-Graz, 1985-2000: from a scientific pilot to a state-wide multimedia radiological information system. In G. J. Blaine, & E. L. Siegel (Eds.), *Proceedings of the SPIE, "Medical Imaging 2000: PACS Design and Evaluation: Engineering and Clinical Issues"* (Vol. 3980, pp. 19–28). San Diego, CA.

Graeber, S. (1997). The impact of workflow management systems on the design of hospital information systems. *The American Medical Informatics Association Annual Fall Symposium*. Retrieved February 12, 2008, from http://www.amia.org/pubs/symposia/D004101.PDF

Grimson, J., Stephens, G., Jung, B., Grimson, W., Berry, D., & Pardon, S. (2001). Sharing health-care records over the Internet. *IEEE Internet Computing, 5*(3), 49–58.

Hasselbring, W., Peterson, R., Smits, M., & Spanjers, R. (2000). *Strategic information management for a Dutch university hospital*. Retrieved February 12, 2008, from http://citeseer.ist.psu.edu/720240.html

Heitmann, K. U. (2002). *XML encoding rules of HL7 v2 messages - v2.xml*. Retrieved February 12, 2008, from http://www.hl7.org/Special/committees/xml/drafts/v2xml1mlbb.pdf

HL7. (2006). *HL7 ANSI-approved standards*. Retrieved February 12, 2008, from http://www.hl7.org/document-center/public/faq/ansi_approved.htm

HL7. (2007). *HL7 EHR system functional model release 1*. Retrieved February 12, 2008, from http://ftp2.ansi.org/download/free_download.asp?document=ANSI%2FHL7+EHR%2C+R1%2D2007

HL7 Canada. (2002). *Guidebook for HL7 development projects in Canada*. Retrieved February 12, 2008, from http://sl.infoway-inforoute.ca/downloads/hl-7can_Guidebook_v1_0.pdf

HL7.org. (2006). *What is HL7?* Retrieved February 12, 2008, from http://www.hl7.org/about/hl7about.htm

Hollingsworth, D. (1995, January 19). Workflow management coalition. *The Workflow Reference Model. Document Number TC00-1003, Issue 1.1*. Retrieved February 12, 2008, from http://www.wfmc.org/standards/docs/tc003v11.pdf

Iacucci, E., Nielsen, P., & Berge, I. (2002, August). Bootstrapping the electronic patient record infrastructure. In K. Bødker, M. Kühn Pedersen, J. Nørbjerg, J. Simonsen, & M. Thanning Vendelø (Eds.), *Proceedings of the 25th Information Systems Research Seminar in Scandinavia (IRIS 25), "New Ways of Working in IS"*. Bautahøj, Denmark. Retrieved February 12, 2008, from http://heim.ifi.uio.no/~pnielsen/PHD/IaNiBe.doc

Interfaceware. (2006). *HL7 message sample.* Retrieved February 12, 2008, from http://www.interfaceware.com/manual/ch-2-3-1-2.html

Jung, B. (2005, June 23–24). DICOM-X - seamless integration of medical images into the EHR. In A. Tsymbal, & P. Cunningham (Eds.), *The 18th IEEE International Symposium on Computer-Based Medical Systems: Computer Based Medical Systems (CBMS)* (pp. 203–207). Trinity College, Dublin, Ireland.

Kennedy, R. L., Seibert, J. A., & Hughes, C. J. (2000, February 15). Legacy system integration using Web technology. In G. J. Blaine, & E. L. Siegel (Eds.), *Proceedings of the SPIE, "Medical Imaging 2000: PACS Design and Evaluation: Engineering and Clinical Issues"* (Vol. 3980, pp. 231–234). San Diego, CA.

Kolesnikov, A. (1997). *EUROMED DICOM WWW gateway*. Retrieved February 12, 2008, from http://cs.joensuu.fi/~koles/projects/wwwdcm/demo.html

Kolesnikov, A., Kelle, O., & Kauranne, T. (1997, May 22). Remote medical analysis with Web technology. In B. Ramin, & P. Karttunen (Eds.), *Proceedings of the Finnish Symposium on Signal Processing - FINSIG'97*, (pp. 87–96). Pori, Finland.

Maier, M. W. (1998). Architecting principles for systems-of-systems. *Systems Engineering, 1*(4), 267–284.

Maier, M. W. (2005). Research challenges for systems-of-systems. *IEEE International Conference on Systems, Man and Cybernetics, 4*, 3149–3154

McCormick, P. (2006). *DICOM and databases*. White Paper. Retrieved February 12, 2008, from http://icmit.mit.edu/dicomdb.html

Miltchenko, M. V., Pianykh, O. S., & D'Antonio, M. G. (2002, December 1–6). Using universal PACS: The globally accessible medical imaging system. *RSNA 2002 Annual Meeting.* Chicago, IL. InfoRAD section. Retrieved February 12, 2008, from http://www.unipacs.com/downloads/RSNA_2002.pdf

National Electrical Manufacturers Association (NEMA). (2004). *Digital imaging and communications in medicine (DICOM) part 1: Introduction and overview*. Report No PS. Retrieved February 12, 2008, from http://medical.nema.org/dicom/2004/04_01PU.PDF

NeoTool. (2007). *The HL7 evolution*. Retrieved February 12, 2008, from http://www.neotool.com/pdf/HL7-Version-3-with-HL7-Version-2-History.pdf

Noumeir, R. (2005). Benefits of the DICOM modality performed procedure step. *Journal of Digital Imaging, 18*(4), 260–269.

Pianykh O. S., Castañeda-Zuniga, W. D., D'Antonio, M. G., & Tyler, J. M. (2001a, August 6–12). Developing an advanced Web-accessible PACS. *International Conference on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet, SSGRR 2001*. L'Aquila, Italy. Retrieved February 12, 2008, from http://www.unipacs.com/downloads/PaperItaly.pdf

Pianykh, O. S., Miltchenko, M. V., & Castaneda-Zuniga, W. D. (2001b, November 24–30). Efficient and reliable Web-based DICOM workstation. *RSNA 2001 Annual Meeting*. Chicago, IL. Scientific poster section. Retrieved February 12, 2008, from http://www.unipacs.com/downloads/Poster.pdf

Ratib, O. M., Ligier, Y., Rosset, A., Staub, J. C., Logean, M., & Girard, C. (2000, February 15). Self-contained off-line media for exchanging medical images using DICOM-compliant standard. In G. J. Blaine, & E. L. Siegel (Eds.), *Proceedings of the SPIE, "Medical Imaging 2000: PACS Design and Evaluation: Engineering and Clinical Issues"* (Vol. 3980, pp. 30–34). San Diego, CA.

Revet, B. (1997). *DICOM cook book for implementations in modalities*. Philips Medical Systems. Free download. Retrieved February 12, 2008, from ftp://ftp.philips.com/pub/pms-3003/DICOM_Information/CookBook.pdf

Tornhola, M. J. (2004). Medical imaging data representation with DICOM-XML. In P. Inchingolo, & R. Pozzi-Mucelli (Eds.), *EuroPACS-MIR 2004 in the Enlarged Europe*. Retrieved February 12, 2008 from http://www.tbs.ts.it/europacs2004/papers/99.pdf

KEY TERMS

Clinical Context Object Workgroup (CCOW): More commonly known as Clinical Context Management, enables the clinical end-user to experience the simplicity of interacting with one system, when in reality he or she may be using multiple independent applications through varying interfaces. Defined by HL7 standards, CCOW ensures secure and consistent access to patient data from varied sources.

Clinical Document Architecture (CDA): Also known as the Patient Record Architecture (PRA), provides an exchange model for clinical documents (such as discharge summaries and progress notes). CDA brings the healthcare industry closer to the realization of an electronic medical record.

Comma Separated Values (CSV): This file format is often used to exchange data between differently similar applications. CSV files have become somewhat obsolete, due to XML data exchange possibilities.

Digital Imaging and Communications in Medicine (DICOM): A standard for the distribution and viewing of medical images, such as CT scans, MRIs, and ultrasound. It describes a file format for the distribution of images and provides access to Storage Media independently from specific media storage formats and file structures.

Health Level Seven (HL7): A standard for exchanging information between medical applications. It is an event driven protocol that defines the format and the content of the messages that applications must use when exchanging data with each another in various circumstances.

Modality Performed Procedure Steps (MPPS): The MPPS service works in concert with the Modality Work List service to provide exam status updates. MPPS are used for two purposes: updates the exam state of patient and forward messages to another application entity, such as a RIS or other work list provider.

Modality Work List: The Modality Work List service provides patient demographic information to imaging modalities in the form of a work list. Modality work lists also allow technologists to sort their workload by time period, specific date or time, patient name or identifier, and modality.

Network Time Protocol/Simple Network Time Protocol (NTP/SNTP): NTP is an Internet protocol used to synchronize the clocks of computers to some time reference. SNTP (Simple Network Time Protocol) is basically also NTP, but lacks some internal algorithms that are not needed for all types of servers.

Picture Archive and Communication System (PACS): A network of computers dedicated to the storage, retrieval, distribution, and presentation of images. A PACS system provides a single point of access for images and their associated data. It also interfaces with existing hospital information systems, such as the hospital information system (HIS) and the radiology information system (RIS).

Service Class Provider (SCP): In DICOM, the relationship between both partners is defined by the Service Class description. The SCP plays the "server" role to perform operations and invoke notifications during an Association. An example of a Storage Service Class Provider would be an image storage device. The SCP role includes enabling the applications to receive DICOM objects locally from devices (image sources, workstations, servers, and so on) through the DICOM protocol.

Service Class User (SCU): In DICOM, the relationship between both partners is defined by the Service Class description. The SCU plays the "client" role to invoke operations and perform notifications during an Association. An example of a Storage Service Class User would be an image acquisition device. In this case, the image acquisition device will create and send a DICOM image by requesting that a Service Class Provider store that image.

Service Object-Pair (SOP): SOP classes identify the capabilities of the specific distributed processing for a certain Service Class. When partners agree to use a SOP class, both sides must ensure they will play their role as described, using the context of the enclosing Service Class. In each SOP Class definition, a single Information Object Definition or IOD is combined with one or more services.

Workflow Management System: A system that completely defines, manages, and executes "workflows" through the execution of software whose order of execution is driven by a computer representation of the workflow logic. The workflow is concerned with the automation of procedures where documents, information, or tasks are passed between participants according to a defined set of rules to achieve, or contribute to, an overall business goal (Hollingsworth, 1995).

Exploring Information Security Risks in Healthcare Systems

Amy Ray
Bentley College, USA

Sue Newell
Bentley College, USA
Warwick Business School, UK

INTRODUCTION

The volume and severity of information security breaches encountered continues to increase as organizations, including healthcare organizations, struggle to identify more effective security policies and procedures. Publicly available guidelines such as GASSP or ISO17799 that are designed to facilitate development of effective security policies and procedures have been criticized for, among other things, inadequate attention to differences in organizational security needs (Baskerville & Siponen, 2002), and for inadequate attention to the social dimensions of security problems (Dhillon & Backhouse, 2001). In this contribution, we argue that the diversity of organizational security needs, as well as the need to recognize the social dimensions to security problems, will continue to grow as companies move away from employing unique, proprietary approaches to software and network development, in favor of adopting standards-based plug-and-play applications, and related standards-based methods and technologies designed to enable interorganizational as well as local systems interoperability.

We use complexity science and adaptive structuration theory to support our arguments that current security management policies and procedures focus on what technologies are used, and on planned systems use to the exclusion of unplanned—but real—emergent use and emergent development of systems. A more holistic approach to security that adapts to emergent systems developments—and most importantly, addresses alternative, emergent uses of systems—is needed, we argue. Throughout the article, we use examples from the healthcare sector to illustrate our points. We do this because Electronic Health Record (EHR) systems that will enable information to be shared across a variety of organizations (local doctors' offices, hospitals, health insurance providers, research organizations, and so on) and users (doctors, administrators, nurses, researchers, and so on) are at the early stages of adoption in many countries, so that much can be gained by starting with an informed view of what can lead to security risks, so that policies and practices are adopted that can protect the information that is being shared.

BACKGROUND

Software development has been described as a "craft" industry, because software applications are developed one at a time, and labor is by far the most significant cost of any development project. Various *standards*—or generally agreed-upon activities, methods, functions, protocols, interfaces, systems, equipment, materials, services, processes and products (De Vries, 2005)—have been introduced and employed in efforts to reduce the labor costs associated with IT projects, especially in terms of standards designed to facilitate creation of Web applications (e.g., TCP, HTML, HTTP, XML, SMTP, UDDI, SOAP).

These standards are generally referred to as Web standards or Web-based standards, and their power to provide interoperability between two or more systems has been established for decades. However, while these standards have benefits, it is important to recognize that using standards has an unintended consequence. More specifically, it can be argued that, as a result of *successful* use of Web-based standards for local systems development and systems integration, overall systems architectures are more complex, ultimately resulting in an environment of greater information security risk. In the next section, we explain why standards-based development and integration increase the overall complexity of the systems architecture, and subsequently consider how this influences emergent use and architectural complexity, and so information security risk.

INFLUENCES ON ELECTRONIC HEALTHCARE RECORD SECURITY

IT Standards-Enabled Planned Systems Development and Complexity

At a local level, using IT standards simplifies the process of connecting one computer to a network of other computers. For example, employing Web standards, countless computers and computing devices around the world are connected, making the Web infinitely multidimensional and nonlinear. However, while Web standards simplify individual systems integration efforts, they potentially increase the complexity of the overall system architecture by enabling connections among heterogeneous systems.

Complex systems are defined as systems that interweave components in such a way that they display variation without being random, and result in a structure that is more than the sum of its parts (Lissack & Roos, 2000). Complexity science research has shown that many highly complex systems—including systems as diverse as the central nervous system, the biosphere, the stock market, telecommunications systems, and human immune system—are not only multidimensional and nonlinear, but are also made up of many selfsimilar components or properties that, in turn, enable development of more complex systems. EHR systems are good examples of such complex systems.

A number of healthcare data standards (e.g., medical code, individual and entity, and transaction processing standards) are now in place for electronic transmission of administrative data as a result of the Health Insurance Portability and Accountability Act of 1996 (HIPAA). Insurance claims data for a patient may be filed on a laptop computer or handheld device in a physician's office, processed by a claims manager from a terminal at an HMO, and otherwise accessed from any authorized device by a user with legitimate (or stolen) authorization. The challenge of mapping the network of authorized access points to HIPAA information grows as the network grows, and the network grows as the application of standards grows. In addition, in order to comply with HIPAA, many healthcare organizations have outsourced claims processing to healthcare clearinghouses (Ivans, 2003), increasing the complexity of the flow of data, as well as increasing the number of individuals with access to patient data. Thus, applying HIPAA transaction standards to systems integration efforts in healthcare simplifies the process of filing an insurance claim, but increases the complexity of the overall network architectures in healthcare.

Web Standards and Emergent Systems Development and Use

In an organizational or interorganizational context, Web-based standards are used to develop information systems applications in support of particular business processes. However, considerable research now exists which illustrates how actual use of IT is often different from the intended use. Users improvise by adding applications to the original system in order to support local practices (Ciborra et al., 2000; Pozzebon & Pinsonneault, 2005), and/or by simply using the system in ways not initially anticipated (Orlikowski, 2000). structuration theory (ST) (Giddens, 1979) and adaptive structuration theory (AST) (DeSanctis & Poole, 2004) are proven as useful frameworks for demonstrating how technologies are adapted and used by individuals within organizations. ST and AST address how social structures are changed via the interaction between a user and a technology, thus helping to explain the emergence and evolution of system use. In this sense, the real nature of a technology and its consequences emerge from the actions of individuals as they engage with and use the technology in practice (Orlikowski, 2000). These consequences can only partially be planned for in advance, and will vary across time and space as context, history, and process impacts the ways in which users develop knowledge through their reflective practice with the technology.

We again turn to healthcare to provide another example of this phenomenon. Since EHR are typically enabled by Web-based—as well as propriety—standards, and are employed by a highly diverse set of users, they are prime candidates for emergent development and use. EHR are designed to support documentation of clinical healthcare. However, physicians, researchers, nurses, insurance company employees, pharmacists, and many others are users of these systems, so it is reasonable to expect that different uses of the systems will emerge across user groups. For example, insurance companies are in the very early stages of adopting pay-for-performance systems to reward physicians based on information collected from EHR about the efficacy and efficiency of care provided to patients. Similarly, researchers are searching for new opportunities to use data available in the standardized, categorized formats enforced by EHR. In fact, EHRs are actually examples of emergent systems' development efforts as they emerged from foundational efforts to create standards for HIPAA (Fedorowicz & Ray, 2004).

The following are examples of simple, emergent uses or developments that can have a dramatic impact on the complexity, and thus manageability, of the overall information systems:

- A nurse adds a wireless network within a department without notifying the IT personnel.
- An individual doctor loads a set of organizational data on a USB drive to take home and analyze on a home computer.
- A sales manager finds that an application designed for one use (repeat orders) is also useful for another purpose (individualized marketing).

The purpose of these examples is to show that system connections and uses are dynamic, rather than to say that they cannot be secured. If uses are known, then effective security policies and procedures may be put in place. However, in an environment where uses are regularly and frequently emerging, existing security policies and procedures may quickly become inadequate, as discussed in the following section.

Architectural Complexity and Information Security Risk

We have presented arguments that standards-enabled systems development and integration methods ultimately increase the complexity of organizational systems' architectures, and support unintended emergent systems development and use. Current security frameworks do not factor in the consequences of emergent systems and system uses, which exacerbates problems noted by other researchers of inadequate attention to differences in organizational security needs (Baskerville & Siponen, 2002) and inadequate attention to the social dimensions of security problems (Dhillon & Backhouse, 2001).

The use of well-understood standards to develop uniquely complex systems creates a triple threat for security managers. First, standards are well-documented and well-understood, which means that any known flaws are typically documented. Documented flaws, for example the ability to exploit knowledge about the IP standard to spoof an IP address, means that hackers—as well as security professionals—have access to this knowledge. Second, the growing complexity of overall systems architectures that result from standards-enabled extension and integration translates into decreasing abilities to understand and manage systems. Since security managers may not even be aware of emergent systems developments and uses, rendering the ability to set effective policies for secure use impossible. Third, the complexity of systems in an environment where systems managers may not be aware of every development and use can contribute to an insecure environment by allowing hackers more points of access to exploit, as well as potentially providing more opportunities to cover their tracks when using unknown access points into systems.

Information security guidance is typically either strategic-level guidance based on organizational goals and primary uses of data, or focused on highly detailed technical issues, such as protecting access points with effective passwords, protecting networks with properly configured firewalls, and appropriate use of encryption. The latter category dominates organizational spending, and using terminology from complexity theory may be referred to as a reductionist approach where security management strategies involve application of solutions to small fractions of a system at a time (e.g., passwords to secure access to an application, firewalls to protect network traffic, and so on). Evidence suggests that reductionism also drives IT security investment (Ernst & Young, 2005). Even strategic security guidance such as HIPAA and ISO17799 make recommendations based on consideration of methods for securing systems components rather than considering the vulnerabilities based on alternative uses of technology, or from considering risks associated with the overall architectural choices supporting the information systems infrastructure. Reductionism is clearly necessary, but it is also increasingly insufficient for effective security management, because, as we have demonstrated in this contribution, it ignores the ways in which systems architectures and uses are dynamic and increasingly complex.

FUTURE TRENDS

In most countries, national health record development efforts are in their early stages, so there is time to consider the full gamut of potential risks, and to design these systems with appropriate measures of protection, both technical and organizational. However, there is little research that has focused on the creation of security policies in contexts that are emergent, rather than static (Baskerville & Siponen, 2002). Instead, most approaches to security, as we have seen, are reductionist in nature. These reductionist approaches are always going to be limited for the reasons discussed above. Instead, what seems to be important is to develop holistic approaches to security, which encompass social, as well as technical, elements, and which is dynamic, so that security issues are considered each time a new use is adopted, even if this use is not formally instigated. This could include, for example, dynamic monitoring for changes in systems architectures, and comprehensive training to not only educate users about risks from intended uses of systems, but to also provide them with methods for considering risks associated with emergent uses of systems.

In considering how this could be approached in an organizational context, it is important to identify who will be responsible for championing this more holistic approach. One new management position in many organizations with complex systems is the Information Systems Architect, whose job responsibilities include approving new systems development projects, and using whatever tools and processes are necessary to remain knowledgeable about the true architecture of the existing information system. For example, the information systems architect might employ network monitoring tools on a regular basis to ensure that rogue wireless networks are identified and appropriately controlled. We suggest here that either the Information Systems Architect also have security management responsibilities, or that key security managers such as Chief Security Officers at least occasionally adopt a practice of considering security risk management from a holistic perspective to complement existing reductionist security policies and practices. We further suggest that much work needs to be done to ensure that security policies are reviewed and updated frequently, and that appropriate steps are adopted for ensuring that security policies and practices are as adaptive as the complex, emergent systems they must protect. A good starting point is to consider developing metapolicies for adaptive development of security policies, and when they are needed (Baskerville & Siponen, 2002). In keeping with our analysis, it is also important to emphasize the principles of security management during training, so that the various users of complex systems such as electronic health record systems will build in their own security practices as the system evolves, even when the emergent applications have inherent security gaps. Metapolicies would be useful for establishing this practice as well. Research exploring the best ways to achieve this holistic understanding and appreciation of the importance of securing data in large complex EHR systems is clearly going to be an important challenge for the future.

CONCLUSION

Our research contributes to the literature by demonstrating the potential domino effect of Web standards-enabled systems integration and development efforts. Standards facilitate rapid development of planned and emergent systems, but can ultimately lead to overall systems architectures that are very complex to manage. When this happens, organizations are at increased risk of security breach. Including security features in individual application design will only be effective to the extent that the interactions between multiple applications in the larger system have all been anticipated during the design process, and that the system is used only in ways that were originally intended. As our contribution suggests, neither of these conditions is very likely. This suggests the need for a more holistic approach to security management.

We have presented numerous examples in this article to support our analysis, but much research is still needed. We believe such research is critically important, given the early stages of Web standards-enabled systems development efforts and the potential for either tremendous gain or tremendous chaos, especially in the context of highly complex interorganizational healthcare systems. This research agenda is not only theoretically important, but also tremendously important from a practical point of view, especially in industries where data processed is of a sensitive nature, such as in healthcare. Here, the development of electronic heath record systems that connect a wide variety of disparate organizations and users is in its infancy. Developing approaches now which can help to secure these systems from security breaches is likely to provide huge cost savings, avoiding, for example, litigation claims when a patient's data has been inappropriately divulged, or simply human misery when an individual is unable to get a job because an unethical employer has managed to extract some private data from a health record. Research which can identify approaches to security management that take into account the problems associated with system complexity, as outlined in this article, would thus be very timely.

REFERENCES

Baskerville, R., & Siponen, M. (2002). An information security metapolicy for emergent organizations. *Logistics Information Management, 15*(5/6), 337–346.

Ciborra, C., et al. (2000). *From control to drift: The dynamics of corporate information infrastructures.* Oxford: Oxford University Press.

DeSanctis, G., & Poole, M. (1994). Capturing the complexity in advanced technology use: Adaptive structuration theory. *Organization Science, 5*(2), 121–145.

De Vries, H. (2005). IT Standards Typology. In K. Jakobs (Ed.), *Advanced topics in IT standards and standardization research*. Hershey, PA: Idea Group Publishing.

Dhillon, G., & Backhouse, J. (2001). Current directions in IS security research: Toward socio-technical perspectives. *Information Systems, 11*(2), 127–154.

Ernst, & Young. (2005). *Global information security survey*. Retrieved February 6, 2006, from http://www.ey.com

Fedorowicz, J., & Ray, A. W. (2004). Impact of HIPAA on the integrity of healthcare information. *Int. J. Healthcare Technology and Management, 6*(2), 142–157.

Giddens, A. (1979). *Central problems in social theory*. London: Macmillan.

Holland, J. (1995). *Hidden order: How adaptation builds complexity*. Addison Wesley.

Ivans. (2003). *Provider survey: Claims processing and connectivity in the emerging HIPAA/CMS environment*. Retrieved February 6, 2008, from http://www.ivans.com

Lissack, M., & Roos, J. (2000). *The next common sense: The e-manager's guide to mastering complexity*. London: Nicholas Brealey.

Orlikowski, W. (2000). Using technology and constituting structures: A practice lens for studying technology in organizations. *Organization Science, 11*(4), 404–428.

Pozzebon, M., & Pinsonneault, A. (2005). Global-local negotiations for implementing configuarable packages: Learning from IT research. *Organization Studies, 26*(9), 1353–1376.

KEY TERMS

Complex Systems: Systems that interweave components in such a way that they display variation without being random, and result in a structure that is more than the sum of its parts.

Complexity Science: The study of complex systems, broadly defined.

Emergent Use: The use of systems in ways not initially anticipated, stimulated by users improvising to add applications to the original system in order to support local practices. Emergent use can be planned or unplanned.

Holistic Information Security Approaches: Security management strategies that encompass social as well as technical elements, and which are dynamic so that security issues are considered each time a new use is adopted, even if this use is not formally instigated.

IT Standards: Generally agreed-upon activities, methods, functions, protocols, interfaces, systems, equipment, materials, services, processes, and products that have been introduced and employed, in efforts to reduce the labor costs associated with IT projects.

Metapolicy: Establish how policies are going to be created in an ongoing way.

Reductionist Information Security Approaches: Security management strategies that apply solutions to small fractions of a system at a time (e.g., passwords to secure access to an application, firewalls to protect network traffic, and so on.).

Fast and Robust Fuzzy C-Means Algorithms for Automated Brain MR Image Segmentation

László Szilágyi
Hungarian Science University of Transylvania, Romania

Sándor M. Szilágyi
Hungarian Science University of Transylvania, Romania

Zoltán Benyó
Budapest University of Technology and Economics, Hungary

INTRODUCTION

By definition, image segmentation represents the partitioning of an image into nonoverlapping, consistent regions, which appear to be homogeneous with respect to some criteria concerning gray level intensity and/or texture.

The fuzzy c-means (FCM) algorithm is one of the most widely used method for data clustering, and probably also for brain image segmentation (Bezdek & Pal., 1991). However, in this latter case, standard FCM is not efficient by itself, as it is unable to deal with that relevant property of images that neighbor pixels are strongly correlated. Ignoring this specificity leads to strong noise sensitivity and several other imaging artifacts.

Recently, several solutions were given to improve the performance of segmentation. Most of them involve using local spatial information: the own gray level of a pixel is not the only information that contributes to its assignment to the chosen cluster. Its neighbors also have their influence while getting a label. Pham and Prince (1999) modified the FCM objective function by including a spatial penalty, enabling the iterative algorithm to estimate spatially smooth membership functions. Ahmed, Yamany, Mohamed, Farag, and Moriarty (2002) introduced a neighborhood averaging additive term into the objective function of FCM, calling the algorithm bias corrected FCM (BCFCM). This approach has its own merits in bias field estimation, but it computes the neighborhood term in every iteration step, giving the algorithm a serious computational load. Moreover, the zero gradient condition at the estimation of the bias term produces a significant amount of misclassifications (Siyal & Yu, 2005). Chuang, Tzeng, Chen, Wu, and Chen (2006) proposed averaging the fuzzy membership function values and reassigning them according to a tradeoff between the original and averaged membership values. This approach can produce accurate clustering if the tradeoff is well adjusted empirically, but it is enormously time consuming. In order to reduce the execution time, Szilágyi, Benyó, Szilágyi, and Adam (2003), and Chen and Zhang (2004) proposed to evaluate the neighborhoods of each pixel as a prefiltering step, and perform FCM afterwards. The averaging and median filters, followed by FCM clustering, are referred to as FCM_S1 and FCM_S2, respectively (Chen & Zhang, 2004). Szilágyi et al. (2003) also pointed out that once having the neighbors evaluated, and thus for each pixel having extracted a one-dimensional feature vector, FCM can be performed on the basis of the gray level histogram, clustering the gray levels instead of the pixels, which significantly reduces the computational load, as the number of gray levels is generally smaller by orders of magnitude. This latter quick approach, combined with an averaging prefilter, is referred to as enhanced FCM (EnFCM) (Cai, Chen, & Zhang, 2007). All BCFCM, FCM_S1, and EnFCM suffer from the presence of a parameter denoted by α, which controls the strength of the averaging effect, balances between the original and averaged image, and whose ideal value unfortunately can be found only experimentally. Another drawback is the fact that averaging and median filtering, besides eliminating salt-and-pepper noises, also blurs relevant edges. Due to these shortcomings, Cai et al. (2007) introduced a new local similarity measure, combining spatial and gray level distances, and applied it as an alternative prefiltering to EnFCM, having this approach named fast generalized FCM (FGFCM). This approach is able to extract local information causing less blur then the averaging or median filter, but still has an experimentally adjusted parameter λ_g, which controls the effect of gray level differences.

Another remarkable approach, proposed by Pham (2003), modifies the objective function of FCM by the means of an edge field, in order to eliminate the filters that produce edge blurring. This method is also significantly time consuming, because the estimation of the edge field, which is performed as an additional step in each iteration, has no analytical solution.

In this article, we propose a novel method for MR brain image segmentation that simultaneously aims high accuracy in image segmentation, low noise sensitivity, and high processing speed.

BACKGROUND

The Standard Fuzzy C-Means Algorithm

The fuzzy c-means algorithm has successful applications in a wide variety of clustering problems. The traditional FCM partitions a set of object data into a number of c clusters based on the minimization of a quadratic objective function. When applied to segment gray level images, FCM clusters the intensity level of all pixels (x_k, $k = 1\ldots n$), which are scalar values. The objective function to be minimized is:

$$J_{FCM} = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m}(x_k - v_i)^2 \quad (1)$$

where $m > 1$ is the fuzzyfication parameter, v_i represents the prototype value of cluster i, $u_{ik} \in [0,1]$ is the fuzzy membership function showing the degree to which pixel k belongs to cluster i. According to the definition of fuzzy sets, for any pixel k, we have $\sum_{i=1}^{c} u_{ik} = 1$. The minimization of the objective function is reached by alternately applying the optimization of J_{FCM} over $\{u_{ik}\}$ with v_i fixed, $i = 1\ldots c$, and the optimization of J_{FCM} over $\{v_i\}$ with u_{ik} fixed, $i = 1\ldots c$, $k = 1\ldots n$ (Hathaway, Bezdek, & Hu, 2000). During each cycle, the optimal values are computed from the zero gradient conditions, and obtained as follows:

$$u_{ik}^{*} = \frac{(x_k - v_i)^{-2/(m-1)}}{\sum_{j=1}^{c}(x_k - v_j)^{-2/(m-1)}} \quad (2)$$

and

$$v_i^{*} = \frac{\sum_{k=1}^{n} u_{ik}^{m} x_k}{\sum_{k=1}^{n} u_{ik}^{m}} \quad (3)$$

After adequate initialization of centroids v_i, (2) and (3) are applied alternately until the norm of the variation of vector **v** is less than a previously set small value ε. FCM has invaluable merits in making optimal clusters, but in image processing it has severe deficiencies. The most important one is the fact that it fails to take into consideration the position of pixels, which is also relevant information in image segmentation. This drawback led to introduction of spatial constraints into fuzzy clustering.

Fuzzy Clustering Using Spatial Constraints

Ahmed et al. (2002) proposed a modification to the objective function of FCM, in order to allow the labeling of a pixel to be influenced by its immediate neighbors. This neighboring effect acts like a regularizer that biases the solution to a piecewise homogeneous labeling. The objective function of BCFCM is:

$$J_{BCFCM} = \sum_{i=1}^{c}\sum_{k=1}^{n}\left[u_{ik}^{m}(x_k - v_i)^2 + \frac{\alpha}{n_k}\sum_{r\in N_k} u_{ik}^{m}(x_r - v_i)^2\right] \quad (4)$$

where x_r represents the gray level of pixels situated in the neighborhood N_k of pixel k, and n_k is the cardinality of N_k. The parameter α controls the intensity of the neighboring effect, and unfortunately its optimal value can be found only experimentally. Having the neighbors computed in every computation cycle, this iterative algorithm performs extremely slowly.

Chen and Zhang (2004) reduced the time complexity of BCFCM, by previously computing the neighboring averaging term or replacing it by a median filtered term, calling these algorithms FCM_S1 and FCM_S2, respectively. These algorithms outperformed BCFCM, at least from the point of view of time complexity.

Szilágyi et al. (2003) proposed a regrouping of the processing steps of BCFCM. In their approach, an averaging filter is applied first, similarly to the neighboring effect of Ahmed et al. (2002):

$$\xi_k = \frac{1}{1+\alpha}\left(x_k + \frac{\alpha}{n_k}\sum_{r \in N_k} x_r \right) \tag{5}$$

This filtering is followed by an accelerated version of FCM clustering. The acceleration is based on the idea, that the number of gray levels is generally much smaller than the number of pixels. In this order, the histogram of the filtered image is computed, and not the pixels, but the gray levels are clustered (Szilágyi et al., 2003), by minimizing the following objective function:

$$J_{EnFCM} = \sum_{i=1}^{c}\sum_{l=1}^{q} h_l u_{il}^m (l - v_i)^2 \tag{6}$$

where h_l denotes the number of pixels with gray level equaling l, and q is the number of gray levels. The optimization formulas in this case will be:

$$u_{il}^* = \frac{(l - v_i)^{-2/(m-1)}}{\sum_{j=1}^{c}(l - v_j)^{-2/(m-1)}} \tag{7}$$

and

$$v_i^* = \frac{\sum_{k=1}^{n} u_{ik}^m x_k}{\sum_{k=1}^{n} u_{ik}^m} \tag{8}$$

EnFCM drastically reduces the computation complexity of BCFCM and its relatives (Cai et al., 2007; Szilágyi et al., 2003). If the averaging prefilter is replaced by a median filter, the segmentation accuracy also improves significantly (Cai et al., 2007; Szilágyi, 2006).

FCM with Spatial and Gray Level Constraints

Based on the disadvantages of the aforementioned methods, but inspired of their merits, Cai et al. (2007) introduced a local (spatial and gray) similarity measure that they used to compute weighting coefficients for an averaging pre-filter. The filtered image is then subject to EnFCM-like histogram-based fast clustering. The similarity between pixels k and r is given by the following formula:

$$S_{kr} = \begin{cases} s_{kr}^{(s)} \cdot s_{kr}^{(g)}, & r \in N_k - \{k\} \\ 0, & r = k \end{cases} \tag{9}$$

where $s_{kr}^{(s)}$ and $s_{kr}^{(g)}$ are the spatial and gray level components, respectively. The spatial term $s_{kr}^{(s)}$ is defined as the L_∞-norm of the distance between pixels k and r. The gray level term is computed as $s_{kr}^{(g)} = \exp(-(x_k - x_r)^2 / (\lambda_g \sigma_k^2))$ where σ_k denotes the average quadratic gray level distance between pixel k and its neighbors. Segmentation results are reported more accurate than in any previous case (Cai et al., 2007).

METHOD

Probably the most relevant problem of all techniques presented above, BCFCM, EnFCM, FCM_S1, and FGFCM, is the fact that they depend on at least one parameter, whose value has to be found experimentally. The parameter α balances the effect of neighboring in case of the former three methods, while λ_g controls the tradeoff between spatial and gray level components in FGFCM.

The zero value in the second row of (9) implies that in FGFCM, the filtered gray level of any pixel is computed as a weighted average of its neighbors. Having renounced to the original intensity of the current pixel, even if it is a reliable, noise-free value, unavoidably produces some extra blur into the filtered image. Accurate segmentation requires this kind of effects to be minimized (Pham, 2003).

We propose a set of modifications to EnFCM/FGFCM in order to improve the accuracy of segmentation, without renouncing to the speed of histogram-based clustering. In other words, we need to define a filter that can extract relevant feature information from the image while applied as a prefiltering step, so that the filtered image can be clustered fast based on its histogram. The proposed method consists of the following steps:

1. As we are looking for the filtered value of pixel k, we need to define a small square or diamond-shape neighborhood N_k around it. Square windows of size 3×3 were used throughout this study.
2. We search for the minimum, maximum, and median gray value within the neighborhood N_k, and we denote them by $\min_k$, $\max_k$ and med_k, respectively.

3. We replace the gray level of the maximum and minimum valued pixel with the median value (if there are more than one maxima or minima, replace them all), unless they are situated in the middle pixel *k*. In this latter case, pixel k remains unchanged, just labeled as unreliable value.
4. Compute the average quadratic gray level difference of the pixels within the neighborhood N_k, using the formula:

$$\sigma_k = \sqrt{\frac{1}{n_k - 1} \sum_{r \in N_k} (x_r - x_k)^2} \qquad (10)$$

5. The filter coefficients will be defined as:

$$C_{kr} = \begin{cases} c_{kr}^{(s)} \cdot c_{kr}^{(g)}, & r \in N_k - \{k\} \\ 1, & (r = k) \wedge (x_k \notin \{\max_k, \min_k\}) \\ 0, & (r = k) \wedge (x_k \in \{\max_k, \min_k\}) \end{cases} \qquad (11)$$

The central pixel k will be ignored if its value was found unreliable, otherwise it gets unitary weight. All other neighbor pixels will have coefficients $C_{kr} \in [0,1]$, depending on their space distance and gray level difference from the central pixel. In case of both terms, higher distance values push the coefficients towards 0.

6. The spatial component $c_{kr}^{(s)}$ is a negative exponential of the Euclidean distance between the two pixels *k* and *r*: $c_{kr}^{(s)} = \exp(-L_2(k,r))$. The gray level term depends on the difference $|x_r - x_k|$, according to a bell-shaped function defined as follows:

$$c_{kr}^{(g)} = \begin{cases} \left[\cos\left(\pi (x_r - x_k)/(8\sigma_k)\right)\right]^2, & |x_r - x_k| \le 4\sigma_k \\ 0, & |x_r - x_k| > 4\sigma_k \end{cases} \qquad (12)$$

7. The extracted feature value for pixel *k*, representing its filtered intensity value, is obtained as a weighted average of its neighbors:

$$\xi_k = \frac{\sum_{r \in N_k} C_{kr} x_r}{\sum_{r \in N_k} C_{kr}} \qquad (13)$$

Algorithm

We can summarize the proposed method as follows:

1. Prefiltering step: for each pixel *k* of the input image, compute the filtered gray level value ξ_k, using (10) – (13).
2. Compute the histogram h_l of the prefiltered image.
3. Initialize v_i with valid gray level values, differing from each other.
4. Compute new fuzzy memberships u_{il}, $i = 1 \ldots c$, l $= 1 \ldots q$, using (7).
5. Compute new cluster prototypes v_i, $i = 1 \ldots c$, using (8).
6. If there is relevant change in the v_i values, go back to step 4. This is tested by comparing any norm of the difference between the new and the old vector **v** with a preset small constant ε.

The algorithm converges quickly; however, the number of necessary iterations depends on ε and on the initial cluster prototype values.

RESULTS AND DISCUSSION

In this section we test and compare the accuracy of four algorithms: BCFCM, EnFCM, FGFCM, and the proposed method, on several synthetic and real images. All the following experiments used 3 × 3 window size for all kinds of filtering.

The proposed filtering technique uses a convolution mask whose coefficients are context dependent, and thus computed for the neighborhood of each pixel. Figure 1 presents the obtained coefficients for two particular cases. Section (a) of Figure 1 shows the case, when the central pixel is not significantly noisy, but some pixels in the neighborhood might be noisy or might belong to a different cluster. Under such circumstances, the upper three pixels having distant gray level compared to the value of the central pixel, receive small weights, and this way they hardly contribute to the filtered value. Section (b) of Figure 1 presents the case of an isolated noisy pixel situated in the middle of a relatively homogeneous window. Even though all computed coefficients are low, the noise is eliminated, resulting a convenient filtered value 111. The migration of weights from the local maximum and minimum towards the median

Figure 1. Filter mask coefficients in case of a reliable pixel intensity value (a), and a noisy one (b). The upper number in each cell is the intensity value, while the lower number represents the obtained coefficient. The arrows show, that the coefficients of extreme intensities are contributed to the median valued pixel.

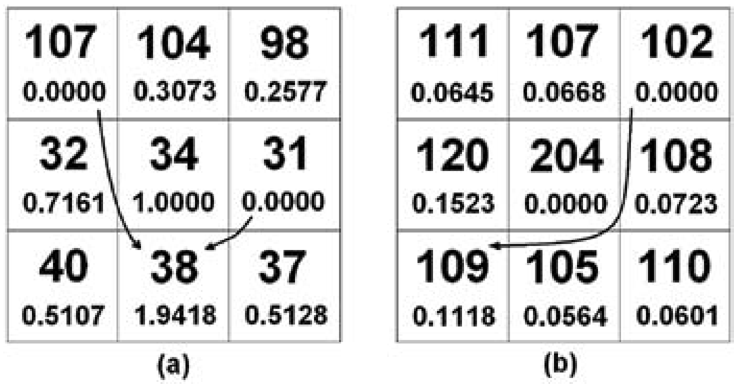
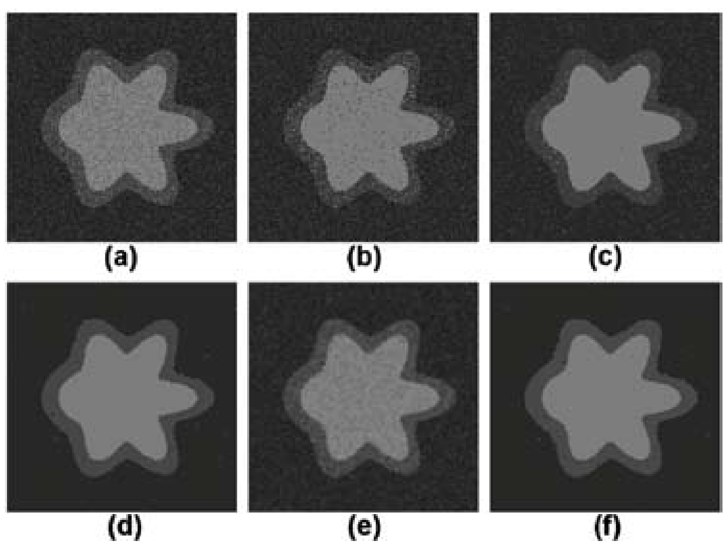


Figure 2. Segmentation results on phantom images: (a) Original, (b) segmented with traditional FCM, (c) segmented using BCFCM, (d) segmented using FGFCM, (e) filtered using the proposed prefiltering, and (f) result of the proposed segmentation.

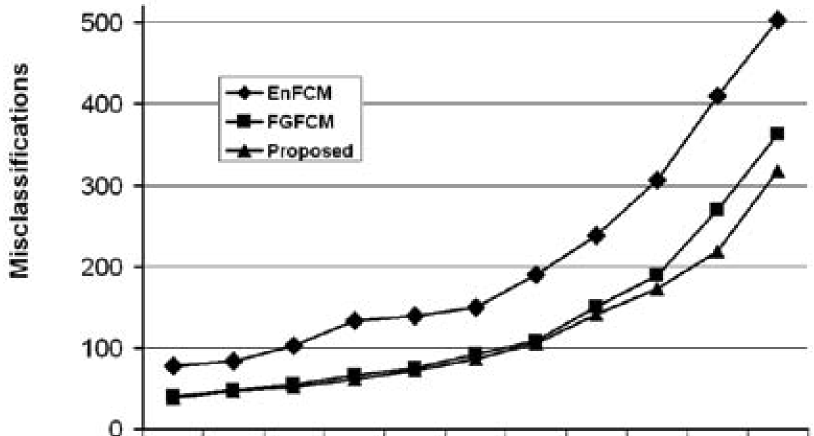


valued pixel, indicated by the arrows, is relevant in the second case and useful in the first.

The noise removal performances were compared using a 256×256-pixel synthetic test image taken from IBSR (Worth) (see section (a) of Figure 2). Figure 2 also shows how much these methods were affected by a high degree mixed noise. Visually, the proposed method achieves best results, slightly over FGFCM, and significantly over all others.

Figure 3 shows the number of misclassifications obtained using three of the presented methods, while segmenting the picture shown in section (a) of Figure 2, corrupted by an increasing amount of mixed noise. Moreover, not only an extra amount of noise is added to the image step by step, but also the original cluster centroids (the base intensities of the clusters) are moved closer and closer. This complex effect is obtained using a weighted sum of three different noisy versions

Figure 3. A comparison of the numbers of misclassifications: Noise level is rising from left to right



	1	2	3	4	5	6	7	8	9	10	11
EnFCM	78	84	103	133	139	150	190	238	306	410	503
FGFCM	40	48	55	66	75	92	108	151	189	270	363
Proposed	38	47	52	61	72	86	106	141	173	219	318

of the same image (all available at IBSR), by changing the weights. Figure 3 reveals that the proposed filter performs best at removing all these kinds of noises. Consequently, the proposed method is suitable for segmenting images corrupted with unknown noises, and in all cases it performs at least as well as his ancestors (Szilágyi, Szilágyi, & Benyó, 2007).

We applied the presented filtering and segmentation techniques to several T1-weighted real MR images. A detailed view, containing numerous segmentations, is presented in Figure 4. The original slice (a) is taken from IBSR. We produced several noisy versions of this slice, by adding salt-and-pepper impulse noise and/or Gaussian noise, at different intensities. Some of these noisy versions are visible in sections (d), (g), (j), and (m) of Figure 4. The filtered versions of the five above mentioned slices are presented in the middle column of Figure 4. The segmentation results of the chosen slice are shown in sections (c), (f), (i), (l), and (o) of Figure 4. From the segmented images we can conclude, that the proposed filtering technique is efficient enough to make proper segmentation of any likely-to-be-real MRI images in clinical practice, at least from the point of view of Gaussian and impulse noises. Table 1 takes into account the behavior of three mentioned segmentation techniques, in case of different noise types and intensities, computed by averaging the misclassifications on 12 different T1-weighted real MR brain slices. The proposed algorithm has lowest misclassification rates in most of the cases.

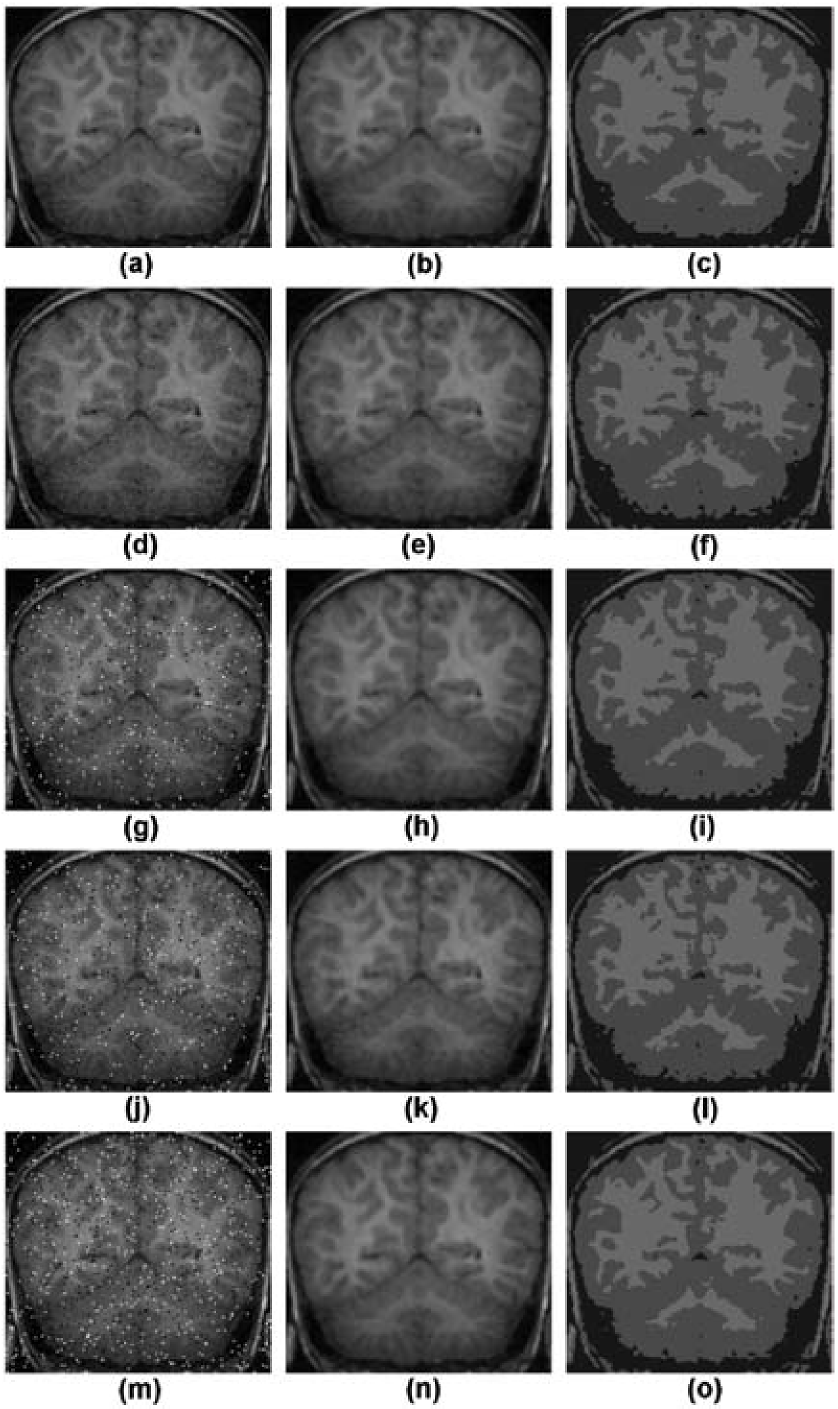
We applied the proposed segmentation method to several complete head MR scans in IBSR. The dimensions of the image stacks were 256 × 256 × 64 voxels. The average total processing time for one stack was around 10 seconds on a 2.4 GHz Pentium 4.

CONCLUSION

We have developed a modified FCM algorithm for automatic segmentation of MR brain images. The algorithm was presented as a combination of a complex prefiltering technique and an accelerated FCM clustering performed over the histogram of the filtered image. The prefilter uses both spatial and gray level criteria, in order to achieve efficient removal of Gaussian and impulse noises without significantly blurring the real edges.

Experiments with synthetic phantoms and real MR images show that our proposed technique accurately segments the different tissue classes under serious noise contamination. We compared our results with other recently reported methods. Test results revealed that our approach outperformed these methods in many aspects, especially in the accuracy of segmentation and processing time.

Figure 4. Filtering and segmentation results on real T1-weighted MR brain images, corrupted with different kinds and levels of artificial noise. Each row contains an original or noise-corrupted brain slice on the left side, the filtered version (using the proposed method) in the middle, and the segmented version on the right side. Row (a)-(c) comes from record number 1320_2_43 of IBSR, row (d)-(f) is corrupted with 10% Gaussian noise, while rows (g)-(i), (j)-(l), and (m)-(o) contain mixed noise of 3% impulse + 5% Gaussian, 3% impulse + 10% Gaussian, and 5% impulse + 5% Gaussian, respectively.



F

Table 1. Misclassification rates in case of real MR brain image segmentation

Noise type	EnFCM	FGFCM	Proposed
Original, no extra noise	0.767 %	0.685 %	0.685 %
Gaussian 4 %	1.324 %	1.131 %	1.080 %
Gaussian 8 %	3.180 %	2.518 %	1.489 %
Gaussian 12 %	4.701 %	2.983 %	2.654 %
Impulse 1 %	0.836 %	0.717 %	0.726 %
Impulse 3 %	1.383 %	0.864 %	0.823 %
Impulse 5 %	1.916 %	1.227 %	0.942 %
Impulse 10 %	3.782 %	1.268 %	1.002 %
Impulse 5 % + Gaussian 4 %	2.560 %	1.480 %	1.374 %
Impulse 5 % + Gaussian 8 %	3.626 %	2.013 %	1.967 %
Impulse 5 % + Gaussian 12 %	6.650 %	4.219 %	4.150 %

FUTURE TRENDS

Further works aim at reducing the sensitivity of the proposed technique to intensity nonuniformity noises and at introducing adaptive determination of the optimal number of clusters.

REFERENCES

Ahmed, M. N., Yamany, S. M., Mohamed, N., Farag, A. A., & Moriarty, T. (2002). A modified fuzzy c-means algorithm for bias field estimation and segmentation of MRI data. *IEEE Transactions on Medical Imaging, 21*, 193-199.

Bezdek, J. C., & Pal, S. K. (1991). *Fuzzy models for pattern recognition*. Piscataway, NJ: IEEE Press.

Cai, W., Chen, S., & Zhang, D. Q. (2007). Fast and robust fuzzy c-means algorithms incorporating local information for image segmentation. *Pattern Recognition, 40,* 825-838.

Chen, S., & Zhang, D. Q. (2004). Robust image segmentation using FCM with spatial constraints based on new kernel-induced distance measure. *IEEE Transactions on System, Man and Cybernetics, Part B, 34*, 1907-1916.

Chuang, K. S., Tzeng, H. L., Chen, S., Wu, J., & Chen, T. J. (2006). Fuzzy c-means clustering with spatial information for image segmentation. *Computerized Medical Imaging and Graphics*, *30*, 9-15.

Hathaway, R. J., Bezdek, J. C., & Hu, Y. (2000). Generalized fuzzy c-means clustering strategies using L_p norm distances. *IEEE Transactions on Fuzzy Systems*, *8*, 576-582.

Pham, D. L., & Prince, J. L.(1999). Adaptive fuzzy segmentation of magnetic resonance images. *IEEE Transactions on Medical Imaging*, *18*, 737-52.

Pham, D. L. (2003). Unsupervised tissue classification in medical images using edge-adaptive clustering. In *25th Annual International Conference of the IEEE EMBS*, Cancún, Mexico (pp. 634-637).

Siyal, M. Y., & Yu, L. (2005). An intelligent modified fuzzy c-means based algorithm for bias field estimation and segmentation of brain MRI. *Pattern Recognition Letters*, *26*, 2052-2062.

Szilágyi, L., Benyó, Z., Szilágyi, S. M., & Adam, H. S. (2003). MR brain image segmentation using an enhanced fuzzy C-means algorithm. In *25th Annual International Conference of the IEEE EMBS*, Cancún, Mexico (pp. 724-726).

Szilágyi, L. (2006) Medical image processing methods for the development of a virtual endoscope. *Periodica Polytechnica*, *Series of Electrical Engineering*, *50*, 69-78.

Szilágyi, L., Szilágyi, S. M., & Benyó, Z. (2007). Efficient feature extraction for fast segmentation of MR brain images. *Lecture Notes in Computer Science*, *4522*, 611-620.

Worth, A. (2001). Internet Brain Segmentation Repository. Retrieved February 1, 2008, from http://www.cma.mgh.harvard.edu/ibsr

KEY TERMS

Data Clustering: Partitioning data sets into subsets whose elements share common or similar properties.

Gaussian Noise: Noise having Gaussian distribution.

Image Segmentation: Partitioning images into regions which are homogeneous with respect to some criteria that concern color, intensity, or texture.

Low-Pass Filter or Averaging Filter: Linear smoothing operator that replaces the intensity of each pixel with a weighted average intensity of its neighbor pixels.

Median Filter: A non-linear filtering technique which replaces the intensity of each pixel with the median intensity value found in a symmetrical neighborhood containing an odd number of pixels. It is usually applied to reduce salt-and-pepper noise.

Salt-and-Pepper Noise or Impulse Noise: Scattered black and white pixels replacing actual value.

Feature Selection in Pathology Detection Using Hybrid Multidimensional Analysis

Edilson Delgado-Trejos
National University of Colombia, Colombia

Germán Castellanos
National University of Colombia, Colombia

Luis Gonzalo Sánchez
National University of Colombia, Colombia

Julio F. Suárez
National University of Colombia, Colombia

INTRODUCTION

The basic problem of pathology detection is to classify a given observation in one of *m* known classes. A set of features presumably contains enough information to distinguish among the classes. Traditional statistical methods break down partly because of the increase in the number of observations, but mostly because of the increase in the number of variables associated with each observation (Castellanos, Daza, Sánchez, Castrillón, & Suárez , 2006).

Performance in training of pattern recognition systems to detect pathologies can be increased, if proper feature extraction is done. Training procedures usually deal with a high number of features; nevertheless a high dimension input space means significant processing time, higher cost in the collection of biosignal records since more observations are needed, and the well known curse of dimensionality phenomena (Lee, Lendasse, & Verleysen, 2003). As a result, the whole training performance declines. In this sense, effective feature selection should be carried out to select those features with higher discriminant capability, keeping or even increasing the accuracy of classification procedures (Yu & Liu, 2004). Literature shows wrapper type procedures of feature selection that search in a considerably smaller number of subsets. Sequential Forward Floating Selection (SFFS) is a common example found in this type of procedure (Webb, 2002). High-dimensional datasets present many mathematical challenges as well as some opportunities, and are bound to give rise to new theoretical developments (Fodor, 2002). Dimensional reduction techniques such as Principal Component Analysis (PCA), which is an orthogonal representation of data, allow us in some way to find a lower dimension in transformed space based on data variance, but this procedure does not take into account information about class separability, therefore PCA is not suitable because the direction of maximum variance does not necessarily correspond to the direction of maximum separability (Hyvarinen, 1999).

Dimensionality reduction procedures perform well on sets of correlated features while variable selection methods perform poorly. These methods fail to pick relevant variables because the score they assign to correlated features is too similar, and none of the variables is strongly preferred over another. Hence, feature selection and dimensionality reduction algorithms have complementary advantages and disadvantages. Dimensionality reduction algorithms thrive on the correlation between variables but fail to select informative features from a set of more complex features. Variable selection algorithms fail when all the features are correlated, but succeed with informative variables (Wolf & Bileschi, 2005).

In this work, we propose a feature selection algorithm with heuristic search that uses Multivariate Analysis of Variance (MANOVA) as the cost function. This technique is put to the test by classifying hypernasal from normal voices of CLP (cleft lip and/or palate) patients. The classification performance, computational time, and reduction ratio are also considered by comparing with an alternate feature selection method based on the unfolding of multivariate analysis into univariate and bivariate analysis. The methodology is effective because it has in mind the statistical and geometrical relevance present in the features, which does not summarize the analysis of the separability among classes, but searches a quality level in signal representation.

BACKGROUND

Dimensionality reduction is the representation of high dimension patterns in a subspace of less dimensionality based on either a linear or nonlinear transformation. This transformation optimizes a specific criterion in the subspace (Cohen, Tian, Zhou, & Huang, 2002). There is a concept responsible for directing the context of reduction called *relevance*, according to a given *cost function*.

Definition 1 (Cost function): Let the initial set of features be $\boldsymbol{\xi} = \{\xi_i : i = 1,\ldots,p\}$, where the estimated values for ξ_i constitute an element of multidimensional representation called observation. The set of observations is composed by $\mathbf{X} = \{\mathbf{x}_\ell \in \mathbb{R}^p : \ell = 1,\ldots,n\}$ contained in $span\{\mathbf{X}\}$; where with linear or nonlinear transformations spaces of representation $\boldsymbol{\mathcal{X}}_j = \mathcal{G}_j\{\mathbf{X}\} = \{\boldsymbol{x}_\ell \in \mathbb{R}^M : \ell = 1,\ldots,n\}$ are obtained. These spaces of representation are composed by subsets of features $\hat{\boldsymbol{\xi}}_j = \{\zeta_i : i = 1,\ldots,M\}$, such that, $M \le p$ y $j \in \mathbb{N}$. Let $\mathbf{k} = \{k_r : r \in \mathbb{N}\}$ be the set of class labels such that each observation $\boldsymbol{x}_\ell$ is assigned one and only one class label. A function that according to an associated metric finds a corresponding value to the evaluated data is called cost function. This function can be described as:

$$f_{\hat{\xi}_j} : \mathbb{N} \times \mathbf{H}_x \to \mathbb{R}$$
$$(\mathbf{k}, \boldsymbol{\mathcal{X}}_j) \mapsto f_{\hat{\xi}_j}(\mathbf{k}, \boldsymbol{\mathcal{X}}_j)$$

where $\mathbf{H}_x \subseteq \mathbb{R}^M$. In this article, we will take $f_{\hat{\xi}_j}(\mathbf{k}, \boldsymbol{\mathcal{X}}_j) \equiv f_{\hat{\xi}_j}(\mathbf{k}, \hat{\boldsymbol{\xi}}_j)$ to aid notation.

Definition 2 (Relevance): Let the observation set be $\boldsymbol{\mathcal{X}} = \{\boldsymbol{x}_\ell \in \mathbb{R}^M : \ell = 1,\ldots,n\}$ composed by M features, $\hat{\boldsymbol{\xi}} = \{\xi_i : i = 1,\ldots,M\}$. Let $\mathbf{k} = \{k_r : r \in \mathbb{N}\}$ be the set of class labels, such that each observation $\boldsymbol{x}_\ell$ is assigned one and only one class label. Let δ be a threshold of significance according to a metric (geometrical, statistical, etc). $\boldsymbol{\mathcal{X}}$ is relevant according to a metric of significance, if there is a cost function, $f_{\hat{\mathbf{i}}}$ with the associated metric of significance, such that, the obtained value for $f_{\hat{\xi}}(\mathbf{k}, \hat{\boldsymbol{\xi}}) > \delta$.

Observation 1 (Feature selection): It is a particular form of reducing dimensionality, when the obtained cost function value, $f_{\hat{\xi}}$, associated with a metric, is involved in the relevance criteria used in a mapping or transformation function and a feature subset is obtained. This feature subset directly corresponds to the initial space ξ and is oriented towards maximizing the representation capability as well as minimizing cost.

In general, given a set of observations where each observation is associated to one and only one class label from a set of labels **k**, the problem of feature selection is to find a subset $\hat{\boldsymbol{\xi}}_i \subseteq \boldsymbol{\xi}$, such that, if the cardinal of $\hat{\boldsymbol{\xi}}_i$ is M, and all the M-cardinal subsets are in $\breve{\boldsymbol{\xi}}$, the subset $\hat{\boldsymbol{\xi}}_i$ is that which optimizes a cost function $f_{\hat{\xi}_i}$, such as (Webb, 2002):

$$f_{\hat{\xi}_i}(\mathbf{k}, \hat{\boldsymbol{\xi}}_i) = \max_{\breve{\xi} \subseteq \xi} f_{\breve{\xi}}(\mathbf{k}, \hat{\boldsymbol{\xi}}_i) \qquad (1)$$

Additionally, feature selection according to the cost function can be carried out in two ways (Yu & Liu, 2004):

1. *Wrapper:* When $f_{\hat{\mathbf{i}}_i}$ uses information of the classification procedure in an attempt to minimize classifier error with validation observations (different to the observations used in training).
2. *Filter:* It is a preprocessing of the data based on the optimization of $f_{\hat{\mathbf{i}}_i}$ with regard to a given metric. Thus, the irrelevant, redundant, and non discriminant variables are discarded. This process is independent of the classification procedure and has advantages such as simple implementation.

Feature selection can be carried out by means of extensive or heuristic search algorithms. Extensive search algorithms are supposed to obtain an optimal feature subset after searching the input training space thoroughly, being computationally expensive (Dash & Liu, 2003). On the other hand, heuristical algorithms which are based on empirical rules can reduce the computational complexity, even though the final subset may not be optimal but enough for the classification purpose. Such methods require some stopping criteria (cost function) (Webb, 2002).

F

DIMENSIONAL REDUCTION BY HEURISTIC MULTIVARIATE ANALYSIS

Given a p-dimensional initial training feature space, a sequential multivariate analysis is done. This is obtained by adjusting a heuristic search method and its cost function providing search direction and subset construction rules.

In general, floating techniques might be considered since they are based on combining forward and backward selection that may provide a better subset optimizing a cost function. For instance, Δp_+ features are added and $\Delta p_- > \Delta p_+$ discarded (Devijver & Kittler, 1982).

Statistical Relevance Measures

Let $\{k = 1,...,L\}$ be the set of patterns, where each one has n observations. Let $\boldsymbol{\xi} \in \mathbb{R}^p$ be the set of features, for which the subset $\hat{\boldsymbol{\xi}}_i = \boldsymbol{\xi} \cap \overline{\xi_i}$ is constructed, being $\overline{\xi_i}$ the complement of ξ_i in $\boldsymbol{\xi}$. A feature ξ_i is assumed to be strongly relevant for the given cost function $f_{\hat{\xi}}$ if and only if,

$$f_{\hat{\xi}}(\mathbf{k},(\xi_i,\hat{\xi}_i)) \neq f_{\hat{\xi}}(\mathbf{k},\hat{\xi}_i) \qquad (2)$$

In this sense, Multivariate Analysis of Variance (MANOVA) can be used as the cost function (2); here the relevance measure is the separability between patterns for a given feature space. This is a hypothesis test about the equality or inequality of the mean values $\boldsymbol{m}_l$ for each one of the class populations. In this case, the model of multiple analysis of statistical relevance for each p-dimensional vector $\boldsymbol{x}_{ij}$

$$\boldsymbol{x}_{kj} = \boldsymbol{m}_k + \varepsilon_{kj}\ , \qquad \boldsymbol{m}_{ki} = \boldsymbol{m} + \alpha_k$$

where j is the observation and k the patterns; $\boldsymbol{m}_k$ are the mean vectors for each pattern and ε_{kj} is the perturbation model, $\boldsymbol{m}$ is the overall mean and a is the class perturbation across the overall mean. The comparison of the mean vectors for the k patterns to find significant differences is carried out through the hypothesis test:

$$H_0 : \boldsymbol{m}_1 = \boldsymbol{m}_2 = \cdots = \boldsymbol{m}_L,$$
$$H_1 : \exists \boldsymbol{m}_l \neq \boldsymbol{m}_m;\ \forall l, m \in 1,\ldots,L$$

MANOVA does not find the features per se; it is only a way for validating a given set of features, which is why this is only a cost function. Consequently, the following approach to constructing and evaluating feature subsets is proposed.

MANOVA Progressive Approach

The Wilks' test is often used for MANOVA, and consists on the likelihood ratio test of H_0 given by:

$$\boldsymbol{\Lambda} = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|} \qquad (3)$$

which is known as Wilks' lambda; **H** being the hypothesis matrix, which can be understood as a measure of dispersion among the mean values of the patterns. While **E** is the error matrix, that relates to a dispersion measure among observations within the patterns. H_0 is rejected if the dispersion among patterns is greater than the dispersion among observations within patterns, and so $\boldsymbol{\Lambda} \in [0,1]$ tends to be zero. On the other hand, Wilks' lambda can be similar to an F-statistic though in an inverse manner. A large F-statistic rejects H_0 (Rencher, 2002).

The following are steps for a forward sequential search algorithm based on multivariate statistical relevance function:

1. Calculate the F-statistic (transformation of Wilks' lambda) for one-feature subsets. From these values, the feature with the largest F-statistic is chosen and its cumulative probability value is computed using the F distribution.
2. Construct 2-dimensional subsets, combining the feature previously chosen in step 1) with the remaining features. Each one of these subsets is evaluated through the Wilks' test, and its respective F-statistic is updated.
3. Select the 2-dimensional subset with the largest F-statistic. Calculate its respective cumulative probability using the F-distribution. This value must exceed the calculated value in step (1) to update the subset. Otherwise, terminate the search.
4. Construct the feature subset adding one feature to the updated subset. These new analysis groups correspond to the subset selected in step (3) and each one remaining feature.

5. Return to step (3) and update the subset using the same criteria. Continue to step (4) and repeat the updating process over and over. The algorithm stops as shown in step (3) when added features do not increase the cumulative probability. So, the final size of the subset is p' where $p' \leq p$. In this manner, we can select those features that together are more discriminant.

Note that if during the calculations $\boldsymbol{\Lambda} \to 0/0$, there is a linear dependency in the current subset. Then its F-statistic is forced to zero and so the evaluated subset is rejected (Johnson & Wichern, 2002).

DIMENSIONAL REDUCTION BY UNFOLDING MULTIVARIATE ANALYSIS

In this case, the multivariate analysis of statistical relevance is unfolded in multiple analyses of smaller dimension. With these kinds of analyses an alternative feature selection methodology is constructed, which connects a 3 block cascade as follows: univariate analysis of separability, bivariate analysis of correlation, and heuristic search using a classifier as the cost function.

Univariate Analysis of Separability

It is important to assure that each one of the features $\xi_i \subset \boldsymbol{\xi}; \forall i = 1, 2, \ldots, p$ have enough discriminant power. A statistical relevance measure is proposed as a hypothesis test based on the significant differences of features between patterns:

- H_0: There is no significant difference on the observations in the ξ_i feature between patterns. Consequently, the overall difference is zero.
- H_1: There is a significant difference on the observations in the ξ_i feature between patterns. Consequently, the overall difference is not zero.

In particular, the hypothesis test is based on the t-Student distribution. Let $\boldsymbol{x}_{k=1,2} \in \boldsymbol{\xi}_i = \{x_j : j = 1, \ldots, n\}$ be the observation vectors for each k pattern, and unknown mean $\tilde{m}_{\boldsymbol{x}_k}$ and variance $\tilde{\sigma}^2_{\boldsymbol{x}_k}$.

From the set of n observations per pattern $\tilde{m}_{\boldsymbol{x}_k}$ and $\tilde{\sigma}^2_{\boldsymbol{x}_k}$, $k = 1, 2$ are estimated. Assuming gaussian distributions for $\boldsymbol{x}_k$, the confidence interval of the respective mean estimations, for a α significance level are given by:

$$\left(\tilde{m}_{\boldsymbol{x}_1} - \tilde{m}_{\boldsymbol{x}_2}\right) - t_{1-\alpha/2}\{n_1 + n_2 - 2\}\tilde{\sigma}_{\Delta\tilde{m}_x} \leq m_{\boldsymbol{x}_1} - m_{\boldsymbol{x}_2} \leq \left(\tilde{m}_{\boldsymbol{x}_1} - \tilde{m}_{\boldsymbol{x}_2}\right) + t_{1-\alpha/2}\{n_1 + n_2 - 2\}\tilde{\sigma}_{\Delta\tilde{m}_{1x}} \quad (4)$$

being

$$\tilde{\sigma}^2_{\Delta\tilde{m}_{1x}} = \frac{n_1\tilde{\sigma}^2_{\boldsymbol{x}_1} + n_2\tilde{\sigma}^2_{\boldsymbol{x}_2}}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

where $t_{1-\alpha/2}\{n_1 + n_2 - 2\}$ represents the percentile value of t-distribution with $n_1 + n_2 - 2$ degrees of freedom (in this case, $n_1 = n_2 = n$). When the confidence interval given by (4) contains zero value, H_0 is not rejected. Otherwise, the alternative hypothesis H_1 is accepted. Thus, it is assumed that there is a significant difference between pattern measures if H_0 is rejected. As result, the dimension of the initial feature set is reduced to $\xi \subset \boldsymbol{\xi} \in \mathbb{R}^{p_1}$, $p_1 \leq p$ (Sharma, 1996).

Bivariate Analysis: Component Analysis

Let $\xi_{1\times p_1}$ be a random vector of features with mean value $E\{\xi\} = \boldsymbol{m}_\xi$, so that the covariance matrix Σ_ξ, of size $p_1 \times p_1$ is:

$$\Sigma_\xi = E\left\{\left(\xi - \boldsymbol{m}_\xi\right)\left(\xi - \boldsymbol{m}_\xi\right)^T\right\}$$

and the correlation matrix P from Σ_ξ is defined as:

$$P(m,q) = \frac{\Sigma_\xi(m,q)}{\sqrt{\Sigma_\xi(m,m)\Sigma_\xi(q,q)}}$$

A null value in the correlation function between the variables ξ_m y ξ_q implies both variables are linearly independent. In the opposite case, the correlation matrix P tends to be singular. Thence, we search for pairs of features that expose larger correlation indexes than \$0.5\$. From each one of these pairs we choose the feature with the smallest overall correlation. Therefore, the new feature subset has dimension p_2, where $p_2 \leq p_1 \leq p$ (Jolliffe, 2002).

Heuristic Search

With the aim of finding a feature subset that minimizes the classification error, a heuristic search is structured. More specifically, we use a forward sequential search algorithm with a cost function based on a Bayesian classifier. Hence, the new subset has dimension p', where $p' \leq p_2 \leq p_1 \leq p$ (Doltsinis, 1999).

EXPERIMENTAL BACKGROUND

A procedure is presented with which the performances of the previously proposed methodologies are evaluated and compared.

Database and Data Preprocessing

The sample contains 90 children recordings (observations) labelled as *normal* and *hypernasal* (45 patients per pattern) and evaluated by specialists. Each recording is composed of five Spanish words: /*coco*/, /*gato*/, /*jugo*/, /*mano*/, and /*papá*/. Signals were acquired under controlled conditions for low ambient noise using a dynamic microphone. The extracted features are based on statistical moments relative to position, scale, and shape per voice parameter (Castellanos et al., 2006). A total of 128 features per word were extracted. Consequently, the initial feature set is 640-dimensional. Data preprocessing is carried out in order to reduce the influence of factors such as systematic acquisition errors, occasional failures of the recording devices, and so forth. Besides, this guaranties the homogeneity of the statistical properties of the training features.

Data preprocessing consists in a hypothesis test (*t*-student) to generate a confidence interval that accepts the data; we discard features when less than 90% of the observations lie within the interval. Kolmogorov-Smirnov test is performed in order to verify normality condition of the data. Those features rejected by this test are transformed using the Box-Cox transform and tested again. Features that are not normal after transformation are rejected.

Both methods proposed (dimensional reduction by heuristic multivariate analysis and dimensional reduction by unfolding multivariate analysis) were compared in three ways: computational time *T*, feature reduction ratio *RR*, and classification accuracy *CA*. To prove the separability of the chosen subsets we used a Bayesian classifier. Eventually, we could compare the results with the extensive search, but having 640 variables the number of subsets is 4.5624e+192.

RESULTS AND DISCUSSION

Results of the comparison between the two methods are shown in Table 1. In Table 1 we can see how the computational time for the first method is longer than the second; nevertheless, these times just illustrate the computational complexity, but this is not a parameter for choosing a method that is performed off-line. Even though the reduction ratio for the first method is a little less than the second, the classification accuracy is higher. It depends on the implementation requirements to choose any of the methods. Accuracy vs. complexity should be the criteria.

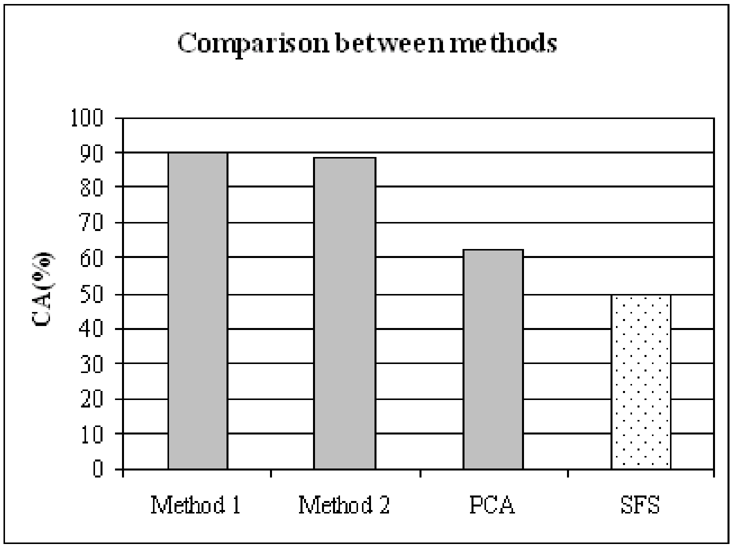
Figure 1 compares dimensional reduction by heuristic multivariate analysis and unfolding multivariate analysis (Method 1 and Method 2 respectively) with traditional methods. We can see that classification improvement is not obtained using only mapping in lower dimension with PCA projection because the direction of maximum separability is not taken into account. In addition to this, SFS algorithm did not converge because the large sets of features generated exhaustive procedures (shown in Figure 1 by the dotted bar).

Both proposed feature selection methods are analyzed for two cases: (a) without data preprocessing and (b) after data preprocessing. From the data preprocess-

Table 1. Comparison of the two methods

Method	T (sec)	RR (%)	CA (%)
Dimensional reduction by heuristic multivariate analysis	19.7	91.26	90
Dimensional reduction by unfolding multivariate analysis	3.81	92.66	88.46

Figure 1. Comparison with traditional methods



ing 36.7% of the original set was discarded by the first test (discard atypical values). The percentage of features that could not be normalized was 25.9%. The results show that in both cases the percentage of reduction is similar although classifier accuracy (CA) increases noticeably when preprocessing is done. This is shown in Table 2 where (—) indicates that the classifier does not converge. Once data preprocessing was done, the initial set was reduced to 47.2%. If we proceed directly to classification after preprocessing the dimension is still too large and the Bayesian classifier does not converge, this is why feature selection is needed.

FUTURE TRENDS

This work allows the study of feature selection to be directed towards different areas of analysis. One of the most important areas of study is the structuring of cost functions, where the efficiency of evolutionary algorithms is proven in the optimization of the solutions. Adjustment of the optimization parameters can be made using regression surfaces, which generate models based on the interaction of the different parameters, thus obtaining the creation of autoadjustable meta-algorithms.

On the other hand, multi-objective MANOVA cost function could be structured such that in a parallel manner it uses measures that not only contain information of separability but also of dependency, redundancy, and irrelevancy of the features.

Table 2. Preprocessing effects

	Method 1 (CA %)	**Method 2 (CA %)**
no preprocessing	–	–
with preprocessing	90	88.46

CONCLUSION

The proposed hybrid methodologies (joint work between multivariate statistical analysis and heuristic search methods for feature subsets), which reduce the initial training space dimension, showed their performance in pathology identification. The methodology is effective because it keeps in mind the statistical and geometrical relevance present in the features, which does not only summarize the analysis of the separability among classes, but also searches a quality level in signal representation. The methodology employed by MANOVA progressive algorithm was the best and it analyzed subsets generated by heuristic principles to avoid exhaustive processes, proposing a decision based on hypothesis testing with low computational cost.

The data preprocessing was a basic stage training system to identify pathologies. In this particular case, the classifier performance does not converge, even employing the best kind of feature selection. In order to comply with all the assumptions of the statistical model imposed, data distribution must be properly analyzed beforehand.

ACKNOWLEDGMENT

This work is within the project titled *"Automatic Hypernasality Detection in Children with Cleft Lip and Palate by Acoustic Analysis"* financed by the Investigations Vice rectory of the National University of Colombia grant 20201004208.

REFERENCES

Castellanos, G., Daza, G., Sánchez, L., Castrillón, O. D., & Suárez, J. (2006). Acoustic speech analysis for hypernasality detection in children. In *IEEE Engineering in Medicine and Biology Society (EMBC06).* New York.

Cohen, I., Tian, Q., Zhou, X. S., & Huang, T. S. (2002). Feature selection and dimensionality reduction using principal feature analysis. In *IEEE international conference on image processing*. Rochester, New York.

Dash, M., & Liu, H. (2003). Consistency-based search in feature selection. *Artificial Intelligence*, *151*, 155-176.

Devijver, P. A., & Kittler, J. (1982). *Pattern recognition: A statistical approach*. Prentice Hall.

Doltsinis, I. (1999). *Stochastic analysis of multivariate systems in computational mechanics and engineering* (1st ed.). Barcelona, Spain: ICNME (International Center for Numerical Methods in Engineering).

Fodor, I. K. (2002, June). *A survey of dimension reduction techniques* (LLNL No. UCRL-ID-148494). Lawrence Livermore National Laboratory.

Hyvarinen, A. (1999). Survey on independent component analysis. *Neural Computing Surveys*, *2*, 94-128.

Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis* (5th ed.). NJ: Prentice Hall.

Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). New York: Spring-Verlag.

Lee, J. A., Lendasse, A., & Verleysen, M.(2003). Nonlinear projection with curvilinear distances: Isomap versus curvilinear distance analysis. *Neurocomputing*, *57*, 49-76.

Rencher, A. C. (2002). *Methods of multivariate analysis* (2nd ed.). New York: Wiley Series in Probability and Statistics.

Sharma, S. (1996). *Applied multivariate techniques*. New York: Wiley.

Webb, A. R. (2002). *Statistical pattern recognition* (2nd ed.). John Wiley & Sons Ltd.

Wolf, L., & Bileschi, S. (2005, June). Combining variable selection with dimensionality reduction. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2*, 801- 806.

Yu, L., & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, *5*, 1205-1224.

KEY TERMS

CLP: Human pathology characterized by cleft lip and/or palate.

Dimensionality Reduction: Data representation in a lower dimension space by linear or nonlinear mapping.

Heuristic Searching: Feature selection by heuristic measure.

Hypothesis Test: Statistical test for accepting or rejecting something by using a distribution.

MANOVA: Hypothesis test based on multivariate analysis of variance.

PCA: Principal Component Analysis. Orthogonal representation based on data variance.

SFFS: Sequential Forward Floating Selection.

An f-MRI Study of an Adaptable EMG Prosthetic Hand with Biofeedback

Alejandro Hernandez Arieta
University of Tokyo, Japan

Ryu Kato
University of Tokyo, Japan

Hiroshi Yokoi
University of Tokyo, Japan

Tamio Arai
University of Tokyo, Japan

Wenwei Yu
University of Chiba, Japan

Rolf Pfeifer
University of Zurich, Switzerland

INTRODUCTION

The mutual adaptation between man-machine opens new possibilities in the development of better user-friendly interfaces that not only adapt to the user's characteristics, but also permits the adaptation of the user to the machine. There are several examples of the use of feedback to improve the man-machine interface. One example is the use of sound to acquire cues in the interaction with the machine (Rauterberg, 1999). These studies show the improvement in the interaction when we increase the number of communication channels between the man and the machine. The problem with sound cues is that need the conscious effort to be recognized. Hunter, Katz, and Davis (2003) show another example of the importance of increasing the communication channels in the interaction between man and machine. In this study, they show how the multiple sensory stimuli contribute to the conscious awareness of the body, and how it can be used to change the abnormal body awareness that occurs after limb amputation. This effect is also known as cortical reorganization, where the brain after losing the stimuli from the amputated limb, due to the cross-modality, received input signals from the adjacent neurons, resulting in what is called "phantom limb."

With the proper stimuli combination, the body image can be changed, allowing for the human body to adapt to external devices. In order to test this hypothesis, we proposed the used of an adaptable, EMG-controlled prosthetic hand with tactile feedback using electrical stimulation. Lotze (1999; Lotze, Grodd, Birbaumer, Erb, Huse, & Flor, 1999) and Maruishi et al. (2004) showed the positive effects in the use of myoelectrical prosthesis to revert cortical reorganization. Giraux, Sirigu, Schneider, and Dubernard (2001) also show the possibility to revert cortical reorganization in the brain. This is possible, due to the combination of the intentionality from the amputee to move the absent limb, which results in muscular movement, and the visual feedback provided by the myoelectric device. We think that if we include even more sensory channels—in this case, tactile feedback—the adaptation to the prosthetic device can be enhanced, resulting in subconscious control of the device.

BACKGROUND

The development of myoelectric prosthetic hands has advanced incredibly since their introduction in 1960 (Katoh, Yokoi, & Arai, 2006). The myoelectrically controlled devices are preferred over the body powered ones, because the latest are restricted in their area of application, and the cosmetic ones do not provide any advantage to their user on his activities (Stein & Walley, 1983). Still, one mayor drawback of the EMG controlled devices is the minimal or nonexistent biofeedback—that is, information on the status of the prosthetic device, hand, or leg to the body. Our body is a multimodal system that uses several channels to obtain the current status of our bodies; if one channel fails, there are still others that help to provide the missing data. The user of a prosthetic hand needs to overcome the lack of tactile and proprioceptive data with visual feedback, which causes to fatigue faster because of the increment of conscious effort to control the hand (Weir, Childress, & Licameli, 1998). These mechanisms need the implementation of a feedback source that enables the user to develop extended physiological proprioception (Simpson, 1974). We find some examples in the application of "tactile feedback" using vibration (Rios Poveda, 2002) or electrical stimulation (Nozomu, Takashi, & Yasunobu, 1998; Shimojo, Suzuki, Namiki, Saito, Kunimoto, Makino, Ogawa, Ishikawa, & Mabuchi k, 2003; Yoshida, Sasaki, & Nakayama, 2002). On the other hand, in the man-machine interfaces studies, we find haptic interfaces that provide tactile feedback. Their direct application to prosthetics is limited, due to the fact that all these researches focus on the sensory substitution using the finger tips (Samuel, 2002). Regrettably, those cannot be applied to prosthetic devices where the user presents partial or complete loss of the arm, which are our interest in this study. Therefore, we need to find a different way to provide with sensorial information to the human body. It is been demonstrated that the brain works with correlative information; therefore, when provided with simultaneous stimuli, the brain can associate the stimuli into a unique event (Carrie Armel & Ramachandran, 2003). Using this knowledge, we can force the brain into producing new sensations, provided that the stimulus is simultaneous, so the person using a prosthetic hand can have sensorial feedback besides the visual.

INTO THE CROSS-MODAL INTERACTION

The prosthetic system in our laboratory uses an adaptive discrimination method to classify the human intention using the electromyography (EMG) signal from three sensors placed on the forearm of the user (Katoh et al., 2006; Nishikawa, Yu, Maruishi, Watanabe, Yokoi, & Kakazu, 2000). In order to provide the tactile sensation, we use in this study transcutaneous electrical stimulation. This method does not interact directly with the nerves, is easier to develop, does not require complex surgical interventions, and the rejection from the body is minima, also it can be applied during the scanning inside the functional magnetic resonation imaging chamber.

The use of visual or auditory feedback only is not reliable enough to provide with sensory feedback for proper subconscious control of the artificial limb. In order to solve this drawback, we need to provide with tactile or proprioceptive information—that is, a way for the body to interact directly with the environment. In previous studies, the electrical stimulation has been used to provide an on-off signal as sensory feedback, with promising results, but lack of more objective system evaluation. The participants to these experiments shows an improvement in their use of the prosthetic device, and even voice the opinion of "feeling" the robot hand more as part of their bodies. Although these studies present promising results, we still need to understand more the relationship between the man and the machine for better future applications.

We transmit the tactile information from the prosthetic hand to the body using a transcutaneous electrical stimulator developed in our laboratory that works as a transducer between the forces applied over the pressure sensors installed on the finger tips and the palm of the robot hand and the biphasic square signal applied to the body (Hernandez Arieta, Yu, &Yokoi, 2004).

Experimental Setting

Using the prosthetic system described above, we performed a series of experiments to measure the activation on the somatosensory region S1 of the brain to see the effects of the electrical stimulation as tactile feedback for the myoelectric prosthetic hand. In order to measure the effects, we first evaluated the response of the body to the stimulation alone as an unrelated event. Following, we measure the brain activation to the electrical stimulation when is applied as a result of the robot hand touching and object, working along with the intention from the body and the visual feedback.

As stated above, we regulate the stimulation by changing the duty rate of the signal, thus any reference to the intensity in the stimulation will be stated according to the percentage of the duty ratio. The neural transcutaneous electrodes were placed on the upper left arm. This is done in order to avoid the insertion of electrical noise, due to the electrical stimulation on the EMG signals. The distance between electrodes was set to 1 cm. Our control group consisted of two healthy men in their 20s, who do not present any visible physical alteration. We placed the sensors similarly to the amputee.

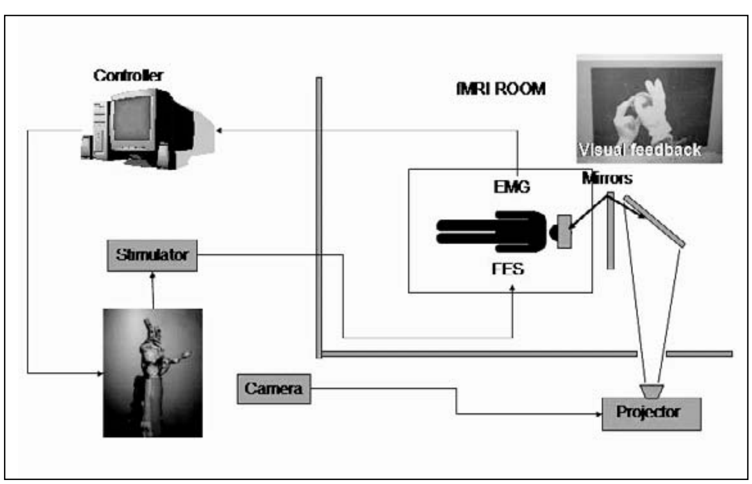
Our test subject for this experiment is a right arm amputee woman in her 50s. The amputation was performed five years previously to this study. The amputation was performed over the wrist level, leaving most part of the forearm intact. The surface EMG sensors were placed over the right forearm, below the elbow joint as shown for the pattern classification process. For this experiment, we trained the classifier with the following motions: fingers flexion/extension, wrist flexion/extension, thumb flexion.

In order to provide with visual feedback, we used a video camera to show the subjects the prosthetic hand movements while inside the f-MRI device. The subjects see the image through a set of mirrors that allow transmit the image from the video camera (Figure 1). This is necessary, because the f-MRI does not allow the inclusion of the robot hand inside the room, due to the strong magnetic force produced by it.

Cerebral activity was measured with f-MRI, using blood oxygen level-dependent contrast (Logothetis, 2002). After automatic shimming, a time course series of 59 volumes was obtained using single-shot gradient-refocused echo-planar imaging (TR = 4000 msec, TE = 60 msec, flip angle = 90 degree, inter-scan interval 8 sec, in-plane resolution 3.44 x 3.44 mm, FOV = 22 cm, contiguous 4-mm slices to cover the entire brain) with a 1.5T MAGNETOM Vision plus MR scanner (Siemens, Erlangen, Germany) using the standard head coil. Head motion was minimized by placing tight—but comfortable—foam padding around the subject's head. The first five volumes of each f-MRI scan were discarded because of non-steady magnetization, with the remaining 54 volumes used for the analysis.

The f-MRI protocol was a block design with one epoch of the task conditions and the rest condition. Each epoch lasted 24 seconds equivalent to three whole-brain f-MRI volume acquisitions. Data were analyzed with Statistical Parametric Mapping software 2000 (http://www.fil.ion.ucl.ac.uk/spm/). The functional magnetic resonance test was set to eight seconds, with a scan time of three seconds, and a rest time of five seconds between scan. 54 scans were acquired for each test. The scans were realigned and transformed to the standard stereotactic space of Talairach using an EPI template. Data were then smoothed in a spatial domain (full width at half-maxim = 8 x 8 x 8 mm) to improve the signal to noise ratio. After specifying the appropriate design matrix, delayed box-car function as a reference waveform, the condition, slow hemodynamic fluctuation unrelated to the task, and subject effects were

Figure 1. Experimental setup; the subject lies inside the f-MRI device with the EMG sensors on the right hand and the neurostimulation electrodes on the left arm; the visual feedback is provided by an array of mirrors that shows the image of the prosthetic hand inside the f-MRI device



estimated according to a general linear model taking temporal smoothness into account. Global normalization was performed with proportional scaling. To test hypotheses about regionally specific condition effects, the estimates were compared by means of linear contrasts of each rest and task period. The resulting set of voxel values for each contrast constituted a statistical parametric map of the t statistic SPM {t}. For analysis of the each session, voxels and clusters of significant voxels were given a threshold of P < 0.005, not corrected for multiple comparisons.

Experimental Procedure

For the experiment related to the hand movement with biofeedback, we first configured the EMG motion classifier (Katoh et al., 2006; Nishikawa et al., 2000) for the subject, that is; we trained the neural network to match the EMG signals for each participant and its corresponding motion in the robot hand. In particular, for this experiment, we used the fingers flexion/extension motion to grab a sphere using the robot hand.

The subject was requested to grab the sphere as soon as she sees it approached by the prosthetic hand in the image produced by the camera inside the f-MRI room. When grabbing the sphere, the pressure sensors placed on the hand were activated, and its pressure signal translated into the transcutaneous electrical stimulation signal of 44% duty rate to ensure that the test subjects perceived the stimulation while inside during the scan acquisition process. We took three scans sessions, with a month between each scan. During this time, the amputee was requested to practice at home with the same set of hand actions using both the EMG classifier and the transcutaneous stimulation.

Results

For the f-MRI data analysis, we used a value of p=0.005 (spm2), resulting in a threshold value of 2.59. Figure 2 shows the evolution of the cortical reorganization and functional specialization in the amputee's brain, due to the myoelectric prosthesis, along with the application of electrical stimulation.

The amputee subject presented activation of the frontal lobe, for the application of electrical stimulation at 44% on the left arm (healthy arm). The f-MRI analysis shows that the electrical stimulation is not detected on the somatosensory area S1 (Figure 3), but is spread on the parietal lobe, in charge of the somatosensory information processing. The test shows that the test subject presents some cortical reorganization.

For the amputee, when grasping the sphere using the robot hand, the activation in the amputee brain increased along the primary motor cortex (M1) related to the right hand. When asked, this person affirmed that there is still the image of the right hand, which is what she used in order to control the prosthetic hand.

Figure 2. Shows the evolution on the amputees brain after the 3 months training period of using the EMG prosthetic hand with biofeedback; at the beginning of the experiment (left) the activation of the brain was spread, showing little specialization of function; this might be caused by the cortical reorganization on the amputee's brain; after the training period (right), the activation in the brain presents a more function specific activation

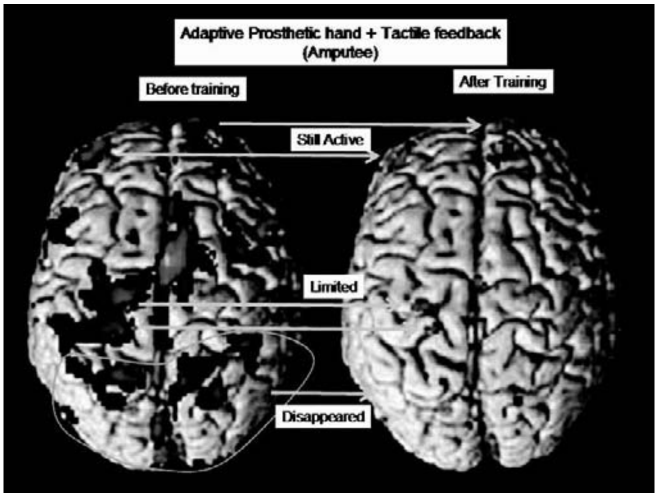


Figure 3. Shows the f-MRI resulting after applying the transcutaneous electrical stimulation on the left arm for the amputee; it is important to notice the activation on the frontal lobe, which identifies the stimulation as new information for the body

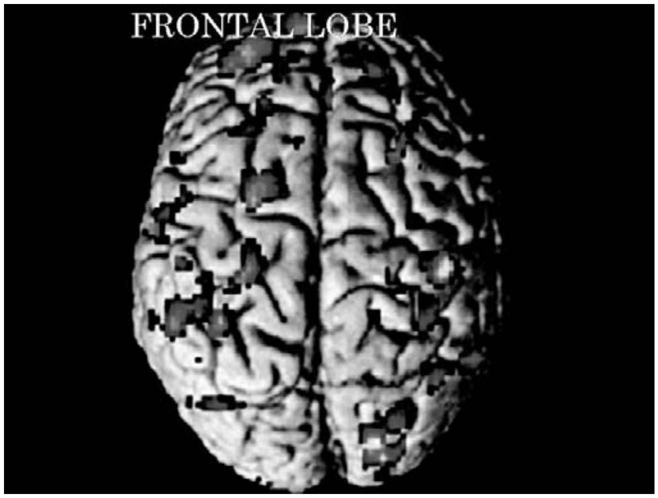


Figure 4 shows an increase in the activation in the primary somatosensory area (S1), principally in the area related to the hand, but also, we see the activation on the area related on the left arm, where the stimulation is actually performed. After the f-MRI scanning, the subjects were questioned on the sensation perceived during the grasping task. The illusion of feeling as if the right hand were touching the object is found in most of the cases.

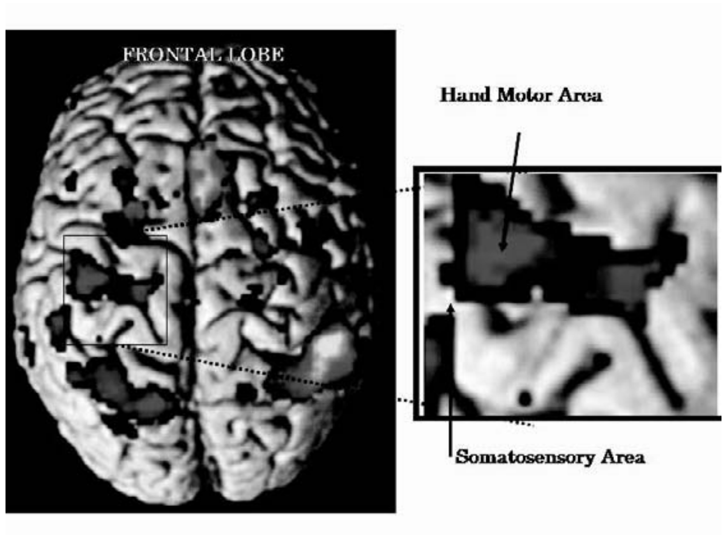
Discussion

Corroborating the findings in neuroscience, the brain does react to the application of electrical stimulation (activation of the frontal lobe), but the activation on the somatosensory region of the left arm is not high enough to appear in the f-MRI image. Therefore, the electrical stimulation alone does not produce an activation of the primary somatosensory area (S1), the area in charge of the processing of sensation in the specific parts of our bodies. When the stimulation is applied in concordance with the action of grabbing an object, we have three channels working altogether: the intention from the subject, the visual feedback, and the stimulus provided by the electrical stimulation. This action is done several times during the scanning process (8 min), which allows the brain to correlate this information as a simultaneous and repetitive event. The f-MRI resulting shows how the brain changes the perception of the electrical stimulation applied on the left arm. Now, the primary somatosensory area (S1) related to the hand presents an activation high enough to be detected in the f-MRI image.

This makes us think that the brain is correlating the multisensorial input as a single event, localizing it in to the right hand (in this case, the prosthetic hand). It is important to notice that the subjects do not interact directly with the object in question, but through the robot hand. They receive only visual feedback through the video display and the stimulation on the left arm.

The application of electrical stimulation does seem to affect the adaptation of the amputee to the use of the prosthetic hand. The application of electrical stimulation as tactile feedback, does allows the body to identify the interaction of the hand with the environment, allowing for a new channel of information to work in benefit of the adaptation to the new “limb.” Also, the results of our experiments over time show the specialization on the activation area in the brain, related to the motor control cortex (MI) and the somatosensory cortex (SI). Another interesting point to discussion is the cross-modality that appears in the brain. In the f-MRI scan, the activation of the somatosensory area is related to the right hand, even though the stimulation is performed on the left arm. This makes us think of a cross-relation in the brain, between the two lobes, allowing adapting the sensory input, and correlating it into a single event. This effect needs to be study with more detail.

Figure 4. Shows the image resulting after the f-MRI scanning for the amputee grabbing a cylinder using the robot hand with electrical stimulation functioning as tactile feedback; the image on the right is the zoom out of the motor and somatosensory area related to the hand and arm; the upper part shows the activation of the motor area in charge of the right hand movement, the lower part shows the reaction from the somatosensory area related to the hand; it is important to notice that the subject does not have the right arm to touch any object



These results show the possibility to use the brain plasticity into the generation of new communication channels with the robotic system. The f-MRI proved a useful tool to measure objectively the changes in the cortical activation, due to the use of the prosthetic system, and allowing a more detailed feedback on the workings of the amputee brain. This allows for a more detailed medical evaluation for the rehabilitation process of an amputee.

The EMG prosthetic hand allows reverting, to some extent, the process of cortical reorganization that occurs when the brain stops receiving the information from the missing limb. This process is even sped up when the use of the EMG prosthetic hand includes feedback to the body. Our experiments with an amputee shows that the electrical stimulation allows an easier adaptation to the prosthetic hand, and actually develop the "illusion" that the robot hand is part of his/her own body, due to the activation on the sensory motor cortex in the brain.

FUTURE TRENDS

As these technologies develop, we can expect an increase on the welfare of those who, for some reason, do not enjoy complete health, with devices helping them to do what others just took for granted. The symbiosis between man and machine will help overcome those barriers that avoid persons with some incapacity enjoy a normal life.

CONCLUSION

In the course of these experiments, we confirmed the importance of the simultaneous stimuli needed for the brain to correlate events, in order to identify them as a single event, opening new channels in the man-machine interaction. The f-MRI is a useful tool to measure objectively the changes in the brain for the interaction with the system proposed. Using the knowledge acquired from previous studies in neuroscience on the brain mechanics, we can now have more efficient man-machine interfaces. Along with continuous use and training, the correlation between visual and sensorial stimuli can be improved, strengthening the illusion of "ownership" of, for example, prosthetic devices. In the future work, we expect to continue measuring the development in the sensorial cortex, due to continuous use of the prosthetic hand with sensorial feedback.

ACKNOWLEDGMENT

We appreciate the collaboration of the Dr. Onhishi Takashi from the Department of Radiology of the National Center Hospital for Mental, Nervous, and Muscular Disorders of Japan for their invaluable support for which this study was possible. This research was partially supported for the project RobotCub: Robotic Open-architecture Technology for Cognition, Understanding and Behavior (Project no. 004370), and the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (B), 2004, No.16360118.

REFERENCES

Carrie Armel, K., & Ramachandran, V. S. (2003). Projecting sensations to external objects: Evidence from skin conductance response. In *Proceedings of R. Soc. B.* (Vol. 270, pp. 1499–1506).

Giraux, P., Sirigu, A., Schneider, F., Dubernard, J. (2001). Cortical reorganization in motor cortex after graft of both hands. *Nature Neuroscience, 4*(7).

Hernandez Arieta, A., Yu, W., Yokoi, H., & et al. (2004). Integration of a multi-D.O.F. In F. Groren et al. Eds.), *Individually Adaptable EMG Prosthetic System with Tactile Feedback. IAS-8* (pp.1013–1021). IOS Press.

Hunter, J., Katz, J., & Davis, K. (2003). The effect of tactile and visual inputs on phantom limb awareness. *Brain, 126*, 579–589.

Katoh, R., Yokoi, H., & Arai, T. (2006). Competitive learning method for robust EMG-to-motion classifier. In *Intelligent Autonomous Systems* (Vol. 9, pp.946–953). IOS Press, ISBN 1-58603-595-9.

Kobrinski, A. E., & et al. (1960). Problems of bioelectric control. In J. F. Coles (Ed.), *Automatic and Remote Control, Proceedings of 1stIFAC International Congress* (Vol. 2, pp.619). Butterworths.

Logothetis, N.K. (2002). The neural basis of the BOLD f-MRI signal. *Phil. Trans. R. Soc. Lond, 357*, 1003–1037.

Lotze, M., Grodd, W., Birbaumer, N., Erb, M., Huse, E., & Flor, H. (1999). Does use of a myoelectric prosthesis prevent cortical reorganization and phantom limb pain? *Nature Neuroscience, 2*(6).

Lotze, M. (1999). Activation of cortical and cerebellar motor areas during executed and imagined hand movements: An f-MRI study. *Journal of Cognitive Neuroscience, 11*, 491–501.

Maruishi, M., et al. (2004). Brain activation during manipulation of the myoelectric prosthetic hand: A functional magnetic resonance imaging study. *Elsevier, Neuroimage, 21*, 1604–1611.

Nishikawa, D., Yu, W., Maruishi, M., Watanabe, I., Yokoi, H., Mano, Y., & Kakazu, Y. (2000). On-line learning based electromyography to forearm motion classifier with motor skill evaluation. *JSME International Journal Series C, 43*(4), 906–915.

Nozomu, H., Takashi, I., & Yasunobu, H. (1998). Sensorial biofeedback using FES. *Biomechanisms Magazine, 12*(1).

Rauterberg, M. (1999). Different effects of auditory feedback in man-machine interfaces. Human Factors in Auditory Warnings. ISBN 0-291-39849-9.

Rios Poveda, A. (2002). *Myoelectric prosthesis with sensorial feedback.* University of New Brunswick's MyoElectric Controls/Powered Prosthetics Symposium.

Samuel, C. (2002). *Behaviour-based control of a robot hand using tactile sensors*. Unpublished master thesis, Center for Autonomous Systems (CAS), Royal Institute of Technology, Sweden.

Shimojo, M., Suzuki, T., Namiki, A., Saito, T., Kunimoto, T., Makino, R., Ogawa, H., Ishikawa, M., & Mabuchi, K. (2003, September, 10). Development of a system for experiencing tactile sensation from a robot hand by electrically stimulating sensory nerve fiber. In *IEEE International Conference on Robotics and Automation* (pp. 1264–1270).

Simpson, D. C., (1974). The choice of control system for the multimovement prosthesis: Extended physiological proprioception (E.P.P.). In P. Herberts, R. Kadefors, R. Magnusson, & I. Petersen (Eds.), *The Control of Upper-Extremity Prostheses and Orthoses*. Springfield: Charles Thomas.

Stein, R. B., & Walley, M. (1983, June). Functional comparison of upper extremity amputees using myoelectric and conventional prostheses. *Arch Phys Med Rehabil., 64*(6), 243–248.

Weir, R. F., Childress, D. S., & Licameli, J. N. (1998). Motor prostheses and the lack of physiologically appropriate feedback: Their function and connections with manipulators, robots and other human-machine systems. In M. Cutkosky, & Srinivasan (Eds.), *Human and Machine Haptics, Section 3.21.* MIT Press.

Yoshida, M., Sasaki, M., Nakayama, N. (2002). *Sensory feedback for biomimetic prosthetic hand*. The 17th Living Body and Physiology Engineering Symposium.

KEY TERMS

Cortical Reorganization: Process believed to occur after a limb amputation due to the lack of stimuli from usual channels. Cause of phantom limbs and phantom pain.

Duty Rate: In electronics is the percentage of time where an square signal presents "On" state.

Electromyography (EMG): A medical technique for evaluating and recording properties of muscles at rest and while contracting. EMG is performed using an instrument called electromyograph, to produce a record called electromyogram. An electromyography detects the electrical potential generated by muscle cells when these cells contract, and also when the cell are at rest.

Echo-Planar Imaging(EPI): MRI Protocol developed by Peter Mansfield to increase the collection speed for T2* weighted images.

Epoch: A reference date. An instant of origin chosen as the date from which time is measured in a computer system.

f-MRI: Functional MRI (fMRI) measures signal changes in the brain that are due to changing neural activity, by a mechanism referred to as BOLD (blood-oxygen-level dependent) effect. Increased neural activity causes an increased demand for oxygen, and the vascular system actually overcompensates for this, increasing the amount of oxygenated hemoglobin (haemoglobin) relative to deoxygenated hemoglobin. Because deoxygenated hemoglobin attenuates the MR signal, the vascular response leads to a signal increase that is related to the neural activity.

Frontal Lobe: Located at the front of each cerebral hemisphere, frontal lobes are positioned in front of the parietal lobes.

Hemodynamic: Literally "blood dynamics," or hemorheology, is the study if the properties and flow of blood. The study of hemodynamics with regard to neural function (the hemodynamic response) is the basis for functional magnetic resonance imaging.

Myoelectric: Also called a motor action potential, is an electrical impulse that produces contraction of muscle fibers in the body. The term is most often used in reference to skeletal muscles that control voluntary movements. Myolectric signals have frequencies ranging from a few Hertz too about 300Hz, and voltages ranging from approximately 10 microvolts to 1 milivolt.

Primary Motor Cortex (M1): Works in association with pre-motor areas to plan an execute movements. M1 contains large neurons known as Betz cells which send long axons down the spinal cord to synapse onto alpha motor neurons which connect to the muscles. Pre-motor areas are involved in planning actions (in concert with the basal ganglia) and refining movements based upon sensory input (this requires the cerebellum).

Primary Somatosensory Area (S1): Brodmann areas 3,1 and 2 comprise the primary somatosensory cortex of the human brain. This area of cortex, as shown by Wilder Oenfield and others, has the pattern of a homunculus. That is, the legs and trunk fold over the midline; the arms and hands are along the middle of the areas shown here; and the face is near the bottom of the figure.

Proprioception: From latin proprius, meaning "one's own" and perception) is the sense of the relative position of neighboring parts of the body. Proprioception is an interception sense that provides feedback solely on the status of the body internally.

Sensory Substitution: Is the principle to transform characteristics of one sensor modality into stimuli of another sensory modality. It is hoped that systems that base on sensory substitution can help handicapped people restore the ability to perceive certain sensory modality.

Transcutaneous Electrical Nerve Stimulation: More commonly referred as TENS unit and pronounced tens, is an electronic device that produces electrical signals used to stimulate nerves through unbroken skin. The name was coined by Dr. Charles Burton. The unit s usually connected to the skin using two or more electrodes.

Fractal Dimension of the EEG in Alzheimer's Disease

Daniel Abásolo
University of Valladolid, Spain

Javier Escudero
University of Valladolid, Spain

Roberto Hornero
University of Valladolid, Spain

Pedro Espino
Hospital Clínico San Carlos, Spain

Carlos Gómez
University of Valladolid, Spain

INTRODUCTION

Alzheimer's disease (AD) is the most frequent cause of dementia in western countries, and is characterized by progressive impairments in cognition and memory, whose course lasts several years prior to death (Jeong, 2004). These clinical features are accompanied by histological changes in the brain, which include widespread cortical atrophy, intracellular deposition of neurofibrillary tangles, and extracellular deposition of senile plaques, particularly in the hippocampus and the cerebral cortex. Although a definite diagnosis is only possible by necropsy, a differential diagnosis with other types of dementia and with major depression should be attempted. Magnetic resonance imaging and computerized tomography can be normal in the early stages of AD, but a diffuse cortical atrophy is the main sign in brain scans. Mental status tests are also useful.

The analysis of the electromagnetic brain activity can provide important information to help in the diagnosis of several mental diseases. The electroencephalogram (EEG) has been used as a tool for diagnosing dementias for several decades. The hallmark of EEG abnormalities in AD patients is a shift of the power spectrum to lower frequencies, and a decrease of coherence among cortical areas (Jeong, 2004).

Recent progress in the theory of nonlinear dynamics has provided new methods for the study of the EEG (Jeong, 2004). Nonlinearity is found in the brain even at the cellular level. Given the highly nonlinear nature of the neuronal interactions, the EEG appears to be an appropriate area for nonlinear time series analysis.

BACKGROUND

There are many studies in which the EEG has been studied with nonlinear time series analysis techniques. These investigations have revealed possible medical applications, since analysis based on nonlinear dynamics yields information unavailable from traditional EEG spectral-band analysis (Pritchard, Duke, Coburn, Moore, Tucker, Jann, & Hostetler, 1994). Moreover, they have given rise to the possibility that the underlying mechanisms of the brain function may be explained by nonlinear dynamics (Röschke, Fell, & Beckmann, 1995; Stam, Jelles, Achtereekte, Rombouts, Slaets, & Keunen, 1995). Particularly, several studies have examined the nonlinear dynamics of the EEG in AD. It has been found that AD patients have lower correlation dimension (D_2) values—a measure of dimensional complexity of the underlying system—than control subjects (Jeong, Chae, Kim, & Han, 2001; Jeong, Kim, & Han, 1998; Pritchard et al., 1994; Stam et al., 1995). The first Lyapunov exponent ($L1$) has also been used to characterize nonlinear behavior. AD patients have significantly lower $L1$ values than controls in almost all EEG channels (Jeong et al., 1998; Jeong et al., 2001a). However, the amount of data required for meaningful results in the computation of D_2 and $L1$ is beyond the experimental possibilities for physiological data (Eckmann & Ruelle, 1992). Thus, different nonlinear techniques are needed to study the EEG background activity.

Mutual information analysis (Jeong, Gore, & Peterson, 2001) has been used to assess information transmission between different cortical areas in AD. Furthermore, some studies have confirmed the decrease of EEG complexity in AD with suitable nonlinear techniques like Lempel-Ziv's complexity (Abásolo, Hornero, Gómez, García, & López, 2006) or multiscale entropy (Escudero, Abásolo, Hornero, Espino, & López, 2006).

As it can be noted, very different nonlinear analysis techniques have been proposed to analyze the EEG background activity in AD patients, although some of them are not suitable to study biomedical signals. In this study, we have evaluated the usefulness of the fractal dimension (*FD*) to characterize the AD patients' EEG background complexity.

FRACTAL DIMENSION OF THE EEG BACKGROUND ACTIVITY

EEG Recording

Ten patients (four men and six women; age = 74.80 ± 3.94 years, mean ± standard deviation (SD)) fulfilling the criteria of probable AD were recruited from the Alzheimer's Patients' Relatives Association of Valladolid (AFAVA) and referred to the University Hospital of Valladolid (Spain), where the EEG was recorded. All of them had undergone a thorough clinical evaluation that included clinical history, physical and neurological examinations, brain scans, and a Mini-Mental State Examination (MMSE), generally accepted as a quick and simple way to evaluate cognitive function (Folstein, Folstein, & McHugh, 1975). The mean MMSE score for the patients was 13.1 ± 5.9 (Mean ± SD).

The control group consisted of 10 age-matched, elderly control subjects without past or present neurological disorders (six men and four women; age = 73.10 ± 6.37 years, mean ± SD). The MMSE score value for all control subjects was 30. All control subjects and all caregivers of the patients gave their informed consent for participation in the current study, which was approved by the local ethics committee.

More than five minutes of data from the 19 scalp loci of the international 10-20 system (electrodes Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T3, T4, T5, T6, Fz, Cz, and Pz) were recorded from each subject using a Profile Study Room 2.3.411 EEG equipment (Oxford Instruments). Sample frequency was 256 Hz, with a 12-bit A-to-D precision. Recordings were made with subjects in a relaxed state, and under the eyes-closed condition, to obtain as many artifact-free EEG data as possible. All EEGs were visually inspected by a specialist physician, and only EEG data free from electrooculographic and movement artifacts, and with minimal electromyographic (EMG) activity were selected. EEGs were organized in five-second artifact-free epochs (1280 points). Furthermore, prior to the *FD* analysis, all recordings were digitally filtered with a band-pass filter with cut-off frequencies at 0.5 Hz and at 40 Hz to remove residual EMG activity.

Fractal Dimension

Fractal geometry is a new language used to describe, model, and analyze complex forms or curves found in nature. Whereas Euclidean shapes have one, or at most a few, characteristic sizes or length scales, fractals, like a coastline, possess no characteristic sizes, and are said to be self-similar and independent of scaling (Accardo, Affinito, Carrozzi, & Bouquet, 1997). A fractal curve in an n-dimensional space has topological dimension n, and a noninteger or fractional dimension called fractal dimension (*FD*). It also possesses the characteristic that each portion of it can be considered a reduced-scale image of the whole for all time scales. If the scaling factor is the same for all time scales, then the curve is said to be self-similar. Many algorithms developed to estimate the *FD* are based on the assumption of self-similarity and independence of scaling (Esteller, Vachtsevanos, Echauz, & Litt, 1999).

FD can be considered a relative measure of signal complexity (Accardo et al., 1997; Katz, 1988). It provides an alternative technique for assessing signal complexity in the time domain, as opposed to the embedding method of assessing this complexity by reconstructing the attractor in the multidimensional phase space, something that is a notable advantage over traditional chaotic techniques, such as D_2 (Accardo et al., 1997). Furthermore, *FD* can characterize different pathophysiological conditions, and it has been particularly useful in the analysis of EEG to characterize neurophysiological states (Esteller et al., 1999; Henderson, Ifeachor, Hudson, Goh, Outram, Wimalaratna, Del Percio, & Vecchio, 2006).

The *FD* also has the advantage of data volume reduction. It is calculated over time in an overlapping sliding window, which greatly reduces the number of data points stored. The exact amount of data depends upon the sliding window size, and on the overlap used for the analysis.

We have estimated the *FD* with Katz's algorithm (Katz, 1988):

$$FD = \frac{\log_{10}(L)}{\log_{10}(d)} \quad (1)$$

L is the length of the curve—the analyzed signal—and *d* its diameter. The length *L* of a signal $x(t) = \{x(0), x(1), ..., x(n)\}$ can be obtained as the sum of the distances between consecutive points:

$$L = \sum_{k=1}^{n} abs\left[x(k-1) - x(k)\right] \quad (2)$$

The diameter *d* is estimated as the distance between the first point of the sequence and the point that provides the farthest distance (Esteller et al., 1999):

$$d = \max\left\{dist\left[x(i), x(0)\right]\right\} \quad (3)$$

The *FD* compares the actual number of units that compose a curve with the minimum number of units required to reproduce a pattern of the same spatial extent. *FD*s computed in this fashion depend on the measurement units used (Esteller et al., 1999). Katz's approach solves this problem by creating a general unit: the average step or average distance between successive points *a*. Normalizing distances in Equation 1 by this average results in:

$$FD = \frac{\log_{10}\left(L/a\right)}{\log_{10}\left(d/a\right)} \quad (4)$$

If we define *n* as the number of steps in the curve, then $L = n \cdot a$, and we can obtain *FD* as:

$$FD = \frac{\log_{10}(n)}{\log_{10}\left(\frac{d}{L}\cdot n\right)} = \frac{\log_{10}(n)}{\log_{10}\left(d/L\right) + \log_{10}(n)} \quad (5)$$

Statistical Analysis

Student's *t*-test was used to evaluate the statistical differences between the *FD* values for AD patients and control subjects. Differences were considered statistically significant if the *p* value was lower than 0.01.

The ability to discriminate AD patients from control subjects at the electrodes where $p < 0.01$ was evaluated using Receiver Operating Characteristic (ROC) curves (Zweig & Campbell, 1993). We define the sensitivity as the rate of patients with a diagnosis of AD who test positive (true positive rate), whereas the specificity represents the fraction of controls correctly recognized (true negative rate). Accuracy is a related parameter that quantifies the total number of subjects precisely classified. The optimum threshold is the cut-off point in which the highest accuracy (minimal false negative and false positive results) is obtained. In addition, leave-one-out cross validation was used to prevent problems like over-fitting and bias.

Results

To estimate *FD*, we defined an overlapping sliding window. We evaluated the *FD* of each EEG epoch with windows of 256 samples (1 s) and a sliding of 64 samples (250 ms, hence an overlap of 750 ms between consecutive windows). For each window, we obtained a *FD* value. In order to obtain a single value for each subject and electrode, the *FD* results were averaged.

FD was estimated for channels Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T3, T4, T5, and T6. The *FD* values (Mean ± SD) for the AD patients and control subjects, and the *p* values of the Student's *t*-tests performed to examine the differences between both groups are summarized in Table 1. As it can be seen, AD patients had lower *FD* values than control subjects at all electrodes, with significant differences between both groups at electrodes T5, P3, P4, O1, and O2 ($p < 0.01$).

Table 1. Average FD values of the EEGs for the AD patients and control subjects for all channels. Significant group differences are marked with an asterisk.

Electrode	Control subjects (Mean ± SD)	**AD** patients (Mean ± SD)	*p*-value
F3	2.1313 ± 0.1334	2.0052 ± 0.0856	0.0215
F4	2.1316 ± 0.1218	2.0264 ± 0.1385	0.0881
F7	2.1673 ± 0.1530	2.0679 ± 0.1349	0.1410
F8	2.1657 ± 0.1358	2.0770 ± 0.1761	0.2230
Fp1	2.1310 ± 0.1276	1.9787 ± 0.1521	0.0260
Fp2	2.1067 ± 0.1662	1.9760 ± 0.1163	0.0565
T3	2.3456 ± 0.1978	2.2166 ± 0.2356	0.2010
T4	2.3223 ± 0.1390	2.2506 ± 0.3275	0.5320
T5*	2.2276 ± 0.1886	1.9904 ± 0.1646	0.0077
T6	2.2302 ± 0.1564	2.0128 ± 0.2012	0.0147
C3	2.1963 ± 0.1261	2.0606 ± 0.1133	0.0210
C4	2.2045 ± 0.1111	2.0863 ± 0.1216	0.0358
P3*	2.1675 ± 0.1191	1.9225 ± 0.1508	0.0008
P4*	2.1838 ± 0.1230	1.9489 ± 0.1543	0.0014
O1*	2.2744 ± 0.1749	2.0035 ± 0.1696	0.0025
O2*	2.2490 ± 0.1865	1.9968 ± 0.1826	0.0068

Finally, we evaluated the ability of the *FD* to discriminate AD patients from control subjects at the electrodes, where significant differences were found using ROC plots and a leave-one-out cross-validation procedure. This technique works by training the models using all available data except for data from one subject; this excluded set is then used to test the performance of the models. The sensitivity, specificity, accuracy, and average optimum thresholds values (*FD* values) obtained are shown in Table II. It can be seen that the accuracy was higher than 75% in 4 of the 5 electrodes, where significant differences between the *FD* values of both groups were found, with a maximum value of 85% at P4. Specificity was 80% at the parietal and occipital electrodes, while sensitivity was 70% at P3, O1, and O2, and 90% at P4. It should be noted that the accuracy was especially poor at T5 (50%).

Discussion

In this pilot study, we analyzed the EEG background activity of 10 control subjects and 10 patients with AD, applying *FD*. *FD* is a nonlinear technique that estimates the complexity of a signal, with increasing values corresponding to increasing complexity. Our results show that *FD* of AD patients' EEG is lower than in control subjects, with significant differences between both groups at electrodes T5, P3, P4, O1, and O2 ($p < 0.01$). These results suggest that the EEG background activity of AD patients is less complex than in a normal brain in the parietal, occipital, and back left temporal regions. The decrease of complexity in AD found with *FD* is in agreement with other studies where the EEG complexity in AD patients was estimated with D_2 (Besthorn, Zerfass, Geiger-Kabisch, Sattel, Daniel, Schreiter-Gasser, & Förstl, 1997; Jeong et al., 2001a; Pritchard et al., 1994; Stam et al., 1995). However, *FD* does not require a large number of samples to be correctly estimated (Accardo et al., 1997). Moreover, the computational cost of estimating *FD* is much lower than in other classical nonlinear techniques derived from chaos theory, given the fact that the reconstruction of the signal attractor in the phase space is not necessary (Accardo et al., 1997). The reduction of complexity in the EEG of AD patients could be explained by a decrease of dynamical complexity of part of the brain. However, the pathophysiological implications of this decrease of EEG complexity are not clear. Among others, three

Table 2. Average sensitivity, specificity and accuracy values obtained with ROC curves and a leave-one-out cross-validation procedure; the average optimum threshold (FD value) that provides these values is also included.

Electrode	Threshold (mean±SD)	Sensitivity (%)	Specificity (%)	Accuracy (%)
T5	2.0876 ± 0.0738	50	50	50
P3	2.0486 ± 0.0274	70	80	75
P4	2.0818 ± 0.0018	90	80	85
O1	2.0567 ± 0.0265	70	80	75
O2	2.0755 ± 0.0047	70	80	75

mechanisms can be responsible for it: neuronal death, a general effect of neurotransmitter deficiency, and loss of connectivity of local neural networks as a result of nerve cell death (Jeong, 2004).

We evaluated the diagnostic accuracy of *FD* with ROC curves and a leave-one-out cross-validation procedure, obtaining accuracy values over 75% for the parietal and occipital electrodes, with a maximum value of 85% at P4. Good accuracies have been reported when classifying AD patients and control subjects with nonlinear techniques. D_2 correctly classified AD patients and controls with an accuracy of 70% (Besthorn et al., 1997). Moreover, the addition of D_2 and a neural net classification procedure to linear methods improved the classification accuracy of AD up to 92% (Pritchard et al., 1994). Furthermore, we obtained accuracies between 72.72% and 90.91% with sample entropy (Abásolo, Hornero, Espino, Álvarez, & Poza, 2006), Lempel-Ziv's complexity (Abásolo et al., 2006b), or multiscale entropy (Escudero et al., 2006) with a set of patients similar to the one considered in this study, but without a leave-one-out cross-validation procedure. Finally, it has been recently reported that an appropriate *FD* measure could achieve 67% sensitivity to probable AD with a specificity of 99.9% (Henderson et al., 2006). However, the comparison of results is not straightforward, due to the different sets of patients and control subjects studied.

Certain limitations of our study merit attention. Firstly, the sample size was small. Hence, to prove the usefulness of *FD* as a diagnostic tool, this approach should be extended on a much larger patient population before any conclusion can be made of its clinical diagnostic value. Moreover, the detected decrease of EEG complexity is not specific to AD. It appears in several physiological and pathological states, including anesthesia (Zhang & Roy, 2001), vascular dementia (Jeong et al., 2001a), schizophrenia (Röschke et al., 1995), and Parkinson's disease (Stam et al., 1995). Thus, although *FD* might be a helpful tool for recognition of AD, further work must be carried out to examine nonlinear EEG activity in other types of dementia.

FUTURE TRENDS

In the last years, nonlinear analysis has been used to study the EEG background activity in AD patients. The use of nonlinear analysis techniques has offered valuable information on cortical dynamics. In fact, nonlinear dynamics suggest that AD can be a dynamical disease characterized by changes in the qualitative dynamics of physiological processes. However, in order to diagnose AD accurately, a large pool of sample data for autopsy-confirmed AD patients, and the age- and sex-matched healthy controls, and other patient groups, are needed to compare their nonlinear measures with the algorithm and the parameters fixed (Jeong, 2004).

Furthermore, the improvement of the accuracy of differential diagnosis and early detection of AD based on multimodal approaches (i.e., the combination of EEG with other structural and functional neuroimaging methods), drug effects on the EEG dynamics in AD and the association of nonlinear dynamics and cognitive performance in AD should be investigated.

CONCLUSION

Although nonlinear EEG analysis cannot yet be applied as a diagnostic tool, our findings show the possibility to analyze the dynamical behavior of the brain in AD patients, and to detect significant differences with *FD*. Our experimental results prove the potential applications of this nonlinear analysis technique in reflecting differences in the complexity of EEG data time series of patients with a diagnosis of AD and control subjects. The decrease of the EEG *FD* found in the parietal, occipital, and left temporal posterior regions in AD patients leads us to think that EEG analysis with *FD* could be a useful tool to increase our insight into brain dysfunction in this disease.

REFERENCES

Abásolo, D., Hornero, R., Espino, P., Álvarez, D., & Poza, J. (2006a). Entropy analysis of the EEG background activity in Alzheimer's disease patients. *Physiological Measurement, 27*(3), 241–253.

Abásolo, D., Hornero, R., Gómez, C., García, M., & López, M. (2006b). Analysis of EEG background activity in Alzheimer's disease patients with Lempel-Ziv complexity and central tendency measure. *Medical Engineering and Physics, 28*(4), 315–322.

Accardo, A., Affinito, M., Carrozzi, M., & Bouquet, F. (1997). Use of the fractal dimension for the analysis of electroencephalographic time series. *Biological Cybernetics, 77*(5), 339–350.

Besthorn, C., Zerfass, R., Geiger-Kabisch, C., Sattel, H., Daniel, S., Schreiter-Gasser, U., & Förstl, H. (1997). Discrimination of Alzheimer's disease and normal aging by EEG data. *Electroencephalography and Clinical Neurophysiology, 103*(2), 241–248.

Eckmann, J. P., & Ruelle, D. (1992). Fundamental limitations for estimating dimensions and Lyapunov exponents in dynamical systems. *Physica D, 56*(2-3), 185–187.

Escudero, J., Abásolo, D., Hornero, R., Espino, P., & López, M. (2006). Analysis of electroencephalograms in Alzheimer's disease patients with multiscale entropy. *Physiological Measurement, 27*(11), 1091–1106.

Esteller, R., Vachtsevanos, G., Echauz, J., & Litt, B. (1999). Fractal dimension characterizes seizure onset in epileptic patients. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'99)* (pp. 2343–2346). Phoenix, AZ.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). Mini-mental state. A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12*(3), 189–198.

Henderson, G., Ifeachor, E., Hudson, N., Goh, C., Outram, N., Wimalaratna, S., Del Percio, C., & Vecchio, F. (2006). Development and assessment of methods for detecting dementia using the human electroencephalogram. *IEEE Transactions on Biomedical Engineering, 53*(8), 1557–1568.

Jeong, J. (2004). EEG dynamics in patients with Alzheimer's disease. *Clinical Neurophysiology, 115*(7), 1490–1505.

Jeong, J., Chae, J. H., Kim, S. Y., & Han, S. H. (2001a). Nonlinear dynamic analysis of the EEG in patients with Alzheimer's disease and vascular dementia. *Journal of Clinical Neurophysiology, 18*(1), 58–67.

Jeong, J., Gore, J. C., & Peterson, B. S. (2001b). Mutual information analysis of the EEG in patients with Alzheimer's disease. *Clinical Neurophysiology, 112*(5), 827–835.

Jeong, J., Kim, S. J., & Han, S. H. (1998). Nonlinear dynamical analysis of the EEG in Alzheimer's disease with optimal embedding dimension. *Electroencephalography and Clinical Neurophysiology, 106*(3), 220–228.

Katz, M. J. (1988). Fractals and the analysis of waveforms. *Computers in Biology & Medicine, 18*(3), 145–156.

Pritchard, W. S., Duke, D. W., Coburn, K. L., Moore, N. C., Tucker, K. A., Jann, M. W., & Hostetler, R. M. (1994). EEG-based neural-net predictive classification of Alzheimer's disease versus control subjects is augmented by nonlinear EEG measures. *Electroencephalography and Clinical Neurophysiology, 91*(2), 118–130.

Röschke, J., Fell, J., & Beckmann, P. (1995). Nonlinear analysis of sleep EEG data in schizophrenia: Calculation of the principal Lyapunov exponent. *Psychiatry Research, 56*(3), 257–269.

Stam, C. J., Jelles, B., Achtereekte, H. A. M., Rombouts, S. A. R. B., Slaets, J. P. J., & Keunen, R. W. M. (1995). Investigation of EEG nonlinearity in dementia and Parkinson's disease. *Electroencephalography and Clinical Neurophysiology, 95*(5), 309–317.

Zhang, X. S., & Roy, R. J. (2001). Derived fuzzy knowledge model for estimating the depth of anesthesia. *IEEE Transactions on Biomedical Engineering, 48*(3), 312–323.

Zweig, M. H, & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry, 39*(4), 561–577.

KEY TERMS

Alzheimer's Disease (AD): A neurodegenerative disease characterized by progressive cognitive deterioration, together with declining activities of daily living and neuropsychiatric symptoms or behavioral changes.

Attractor: In dynamical systems, a set to which the system evolves after a long enough time, after transients have died out.

Correlation Dimension (D_2): A measure of the complexity of a nonlinear system. It characterizes the distribution of points of the attractor in the phase space.

Electroencephalogram (EEG): Neurophysiologic recording of electrical brain activity by positioning electrodes generally on the scalp.

Fractal: A self-similar structure whose geometrical and topographical features are recapitulated in miniature on finer and finer scales.

Fractal Dimension (*FD*): Statistical quantity that gives an indication of how completely a fractal appears to fill space. It can be considered a measure of signal complexity.

Nonlinear Analysis Techniques: Mathematical techniques to analyze nonlinear systems, whose behavior is not expressible as a sum of the behaviors of its descriptors, or signals whose behavior is nonlinear.

Receiver Operating Characteristic (ROC) Curve: A graphical plot of the sensitivity vs. (1–specificity) for a classifier system, as its discrimination threshold is varied.

A Framework for Information Processing in the Diagnosis of Sleep Apnea

Udantha R. Abeyratne
University of Queensland, Australia

INTRODUCTION

Obstructive sleep apnea (OSA) is one of the most common sleep disorders. It is characterized by repetitive obstruction of the upper airways during sleep. The frequency of such events can range up to hundreds of events per sleep-hour. Full closure of the airways is termed *apnea,* and a partial closure is known as *hypopnea*. The number of apnea/hypopnea events per hour is known as the AHI-index, and is used by clinical community as a measure of the severity of OSA.

OSA, when untreated, presents as a major public health concern throughout the world. OSA patients use *health facilities at twice the average rate* (Delaive, Roos, Manfreda, & Kryger, 1998), causing huge pressures on national healthcare systems. OSA is associated with serious complications such as cardiovascular disease, stroke, (Barber & Quan, 2002; Kryger, 2000,), and sexual impotence. It also causes cognitive deficiencies, low IQ in children, fatigue, and accidents. Australian Sleep Association reported (ASA, 1999) that in the state of New South Wales alone 11,000–43,000 traffic accidents per year were attributable to untreated-OSA.

OSA is a highly prevalent disease in the society. An estimated 9% of the women and 24% of the men in the U.S. population of 30 to 60 years was found to be having at least mild OSA (Young, Evans, Finn, & Palta, 1997). In Singapore, about 15% of the *total population* has been estimated to be at risk (Puvanendran & Goh, 1999). In a recent study in India (Udwadia, Doshi, Lonkar, & Singh, 2004), 19.5% of people coming for routine health checks were found to have at least mild OSA.

The full clinical significance of OSA has only recently been understood. Partly as a result of this, the public awareness of the disease is severely lacking. Healthcare systems around the world are largely unprepared to cater to the massive number of OSA patients. This problem is especially severe in the developing world, where OSA diagnostic facilities are rare to find.

BACKGROUND

Definition of Sleep Apnea and Hypopnea

Sleep apnea refers to a cessation of breathing at night, usually temporary in nature. The American Academy of Sleep Medicine Task Force formally defines *apnea* as:

a. Cessation of airflow for a duration ≥10s, or
b. Cessation of airflow for a duration < 10s (for at least one respiratory cycle) with an accompanying drop in blood oxygen saturation by at least 3%.

Hypopnea is defined as a clear decrease (≥50%) in amplitude from base line of a valid measure of breathing (eg., airflow, air pressure) during sleep for a duration ≥10s, *plus* either:

a. An oxygen desaturation of ≥3%, or
b. An EEG-arousal (EEGA) (Flemons & Buysse, 1999).

The average number of obstructive sleep apnea and hypopnea events per hour of sleep, as computed over the total sleep period, is defined as the *Apnea Hypopnea Index (AHI).*

The Current Standards in Apnea/Hypopnea Diagnosis

The current standard of diagnosis of OSA is Polysomnography (PSG). Routine PSG requires that the patients sleep for a night in a hospital *Sleep Laboratory*, under video observation. In a typical PSG session, signals/parameters such as ECG, EEG, EMG, EOG, nasal/oral airflow, respiratory effort, body positions, body movements, and the blood oxygen saturation are carefully monitored. Altogether, a PSG test involves over 15 channels of measurements *requiring physical contact* with the patient.

Drawbacks of PSG and Possible Improvements

At present, the hospital-based PSG test is the definitive method of diagnosis of the disease. However, it has the following drawbacks, particularly when employment as a community-screening tool is considered:

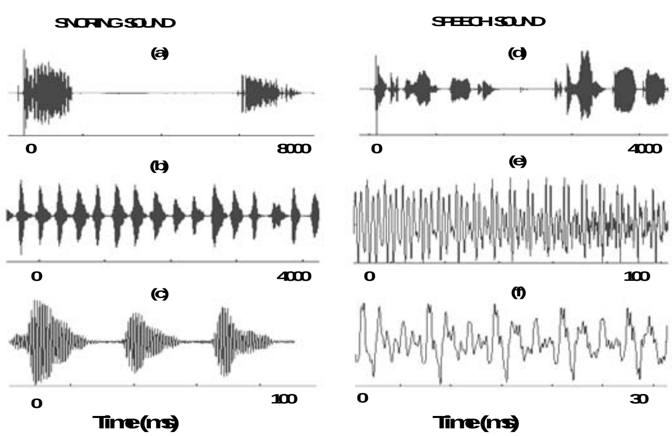
1. *Poor data integrity is a common problem* in routine PSG tests. Even when the test is done in the hospital, it is common to see cases of data loss (or quality deterioration), due to various reasons (eg., improper sensor contact due to electrodes/sensors coming loose, SpO2 sensor falling off, and measurement problems, such as movement artifacts).
2. *PSG interpretation is a tedious task,* due to the size of the data gathered, complexity of the signals, and measurement problems such as data loss.
3. *PSG requires contact instrumentation*; channels such as EEG/ECG/EOG/EMG require Galvanic contact with the patient. It is especially unsuited for pediatric use.
4. *PSG is not suitable for mass screening* of the population. A trained medical technician is required to connect the patient to the PSG equipment, and the patient needs to be monitored overnight to avoid incurring data losses.
5. *PSG is expensive;* this is another factor working against mass screening uses.
6. *AHI index* (and a variant of it known as the respiratory disturbance index, RDI) is used as the golden clinical measure on the severity of apnea. However, AHI (or RDI) does not always correlate strongly with the symptoms of apnea as experienced by patients.

There is an enormous clinical need for a simplified diagnostic instrument capable of convenient and reliable diagnosis/screening of OSA at a home setting (Flemons, 2003). Similarly, hospital-based, full PSG testing requires better measures to characterize the disease. This article explores possible solutions to both of these problems.

There has been a flurry of recent activities at developing technology to address the issue of home screening of OSA. Four different classes of OSA monitors are under development (Flemons, 2003, and references therein). These devices varied from two-channel (eg., airflow and oximetry) systems (designated Type-IV), to miniaturized full-PSG (Type-I) units. Their major drawbacks are:

- Existing take-home devices have *at least* one sensor which requires physical contact. This makes them difficult to use by untrained persons, and cumbersome to use on pediatric populations. TYPE-IV systems, with the smallest number of sensors, suffers from the fact that oxymetry identifies oxygen saturation in blood only as a surrogate for OSA, and the absence of significant desaturation does not mean the absence of the disease (Flemons, 2003).

Figure 1. Similarities and differences between speech and snoring: (a),(b),(c) snoring at different time scales; (d),(e),(f) speech at different time scales



- Presence of a medical technologist is still required, if acceptable sensitivity/specificity performance is required. High rates of data loss (up to 20% loss at home compared with 5% at sleep lab) (Flemons, 2003) (nine) have been reported when a medical technologist is not in attendance. Unattended systems have not led to high enough sensitivity/specificity levels to be used in a routine home monitoring exercise (Flemons, 2003; Portier, 2000).
- Type-I and II devices use channel counts from 7–18 and are difficult to use by an untrained person. The quality and the loss of data is a serious problem.

In this article, we present an instrumentation and signal processing framework addressing current problems in OSA diagnosis.

METHODS

A Framework for the Noncontact Diagnosis of OSA

Snoring almost always accompanies OSA, and is universally recognized as its earliest symptom (Hoffstein, 2000; Kryger, 2000; Puvanendran & Goh, 1999). Logic dictates that it should be providing us with the earliest opportunity to diagnose the disease. At present, however, quantitative analysis of snore sounds is not a practice in OSA detection. The vast potential of using snoring in the noninvasive diagnosis of OSA remains unused.

In this article, we argue that the human speech and snore sounds share many similarities (see Figure 1), and biological "wetware" used for the generation processes. The upper airway acts as an acoustic filter during the production of snoring sounds, just as the vocal tract does in the case of speech sounds. Episodes of OSA, by definition, are associated with partial or full collapse of upper airways. As such, changes to the upper airways brought about by this collapse should be embedded in snoring signals.

Figure 2 shows a typical snore sound recorded from a patient undergoing routine PSG testing at the sleep diagnostic laboratory. The data acquisition system used a sampling rate of 44.1k samples/s, and a bit resolution of 16bits/sample. The system had a 3dB bandwidth of 20.8kHz. A typical hospital based routine PSG test runs for up to eight hours, and it is quite common to observe up to 8,000 events of snores within a recorded sound data.

Snore Sounds: A Working Definition

One of the major problems towards automation is that *there is no objective definition* of what a "snore" is (Hoffstein, 2000). Recently, we proposed (Abeyratne, Wakwella, & Hukins, 2005) an objective definition for "snoring" *independent of the sound intensity*. It is based on the observation that sounds perceived as "snores" by humans are characterized by repetitive packets of energy that are responsible for creating the vibratory sound peculiar to snores (Figure 2, bottom). We call the distance between such packets as the "*pitch*" of snoring. A snoring episode (SE) will consist of a segment with pitch ("voiced-segment"), and possibly "silence" and segments without pitch ("unvoiced segment"). An inspiration-expiration cycle without any segment associated with a pitch is termed as a pure breathing episode (PB).

Mathematical Framework for Sound Analysis

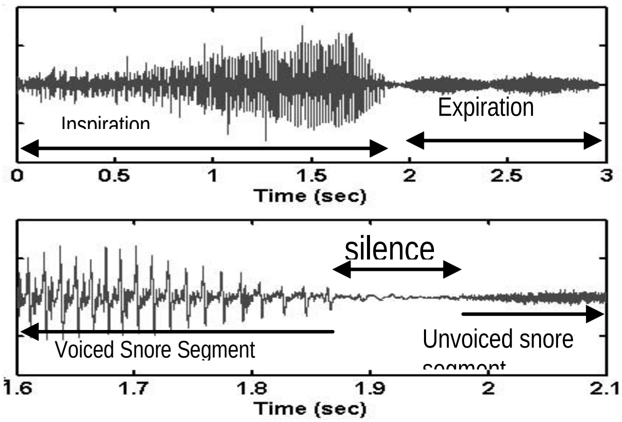
Snoring originates from acoustical energy released by the vibratory motions of upper airway site(s) during sleep. The airway cavity acts as an acoustical filter, and modifies the source sounds to produce snoring sounds we hear. Inspired by the source/vocal-tract system for human speech synthesis, we model a block of recorded sound *s(n)* as a convolution between: (i) "source signal *x(n)*" representing the source acoustical energy, and (ii) "Total Airway Response (TAR) *h(n)*," which captures the acoustical features of the upper airways as given by:

$$s(n) = h(n) * x(n) + b(n) \qquad (1)$$

The symbol *b(n)* denotes background activities, x(n) is the source excitation, "*" denotes convolution, and *h(n)* is the *TAR* function. The term *b(n)* is due to a range of independent reasons and hence considered a Gaussian distribution independent of *x(n)*.

The nature of *x(n)* depends on the nature of the sound segment under consideration. For the case of a segment without any pitch (i.e, either a PB or unvoiced-snoring segment), we model the source excitation, *x(n)* as

Figure 2. (Top) A snore episode (SE); (bottom) different segments in SE



a white random process. The *x(n)* for *voiced snoring segments* is modeled by a pseudo-periodic pulse-train, drawing from techniques used in speech analysis.

The *TAR* function *h(n)* is modeled in this article as a *mixed-phase system* considering that in OSA, multiple-source locations with temporal relations between eachother can be found, and thus, phase-information cannot be neglected in general.

It is our hypothesis that the state of the upper airways can be characterized by the pair $\zeta = \{g\{x(n)\}, f\{h(n)\}\}$, where *g* and *f* represent the operation of extracting features out of the *x(n)* and *h(n)*.

We developed a snore segmentation algorithm (Abeyratne et al., 2005), which takes in full night sound data, and separates into snoring and pure-breathing episodes. It then further divides snore episodes into sub-categories "voiced," "unvoiced," and "silence" sections. The category "voiced" will then be further analyzed to estimate the pitch associated with each segment.

Consider an arbitrary *j*-th Snoring Episode (SE) in the sound recordings. Divide the voiced segment $s_{sv,j}$ of the *j*-th Snoring Episode into L_j number of data blocks $\{B_{jk}\}$, $k = 1\ldots L_j$, each of length *N*. Thus, at the output of the pitch-detector, each data block in the set $\{B_{jk}\}$ is associated with a pitch period μ_{jk}. We term the series $\{\mu_{jk}\}$, k=1,2, .., L_j as the *Intra-snore Pitch Series* for the *j*-th Snoring Episode.

We consider the structure of the *Intra-snore Pitch Series*, and show that it is characterized by discontinuities, which can be used in the diagnosis of OSA. We propose a new measure called *ISPJ-Probability* to capture intra-snore jumps in pitch, via *Definition-1* and *Definition-2* given below:

- ***Definition-I:*** Suppose that in the *j*-th arbitrary snoring episode $s_{sv,j}$ there are at least q ($< N_j + 1$) data blocks in the set $\{B_{jk}\}$, $k = 1\ldots N_j$ with pitch periods μ_{jk} greater than a pitch threshold γ. Then the entire Snoring Episode *j* is labeled as having the feature ISPJ at level (q, γ). We call the quantity q as the "jump multiplicity."
- ***Definition II:*** Define a quantity ISPJ-probability $Pq\gamma(r_D)$ at level (q, γ) for the signal s(n) of a length D as: $Pq\gamma(r_D) = 100\ n_{q\gamma}(r_D)/r_D\%$, where r_D is the total number of snore episodes contained within the data of length D. The quantity $n_{q\gamma}(r_D)$ is the number of episodes within D that were labeled as having the feature ISPJ according to Definition-I.

The snore sound based *test statistic* proposed in this article is $Pq\gamma(r_D)$, with the corresponding decision threshold symbolized by P_{th}. If $Pq\gamma(r_D) > P_{th}$, the *snore-based test* is positive for OSA, and vice versa.

For each subject in the database, a set **S** of *ISPJ-Probabilities* is defined by:

$$S = \{Pq\gamma(r_D) \mid q = 1,2,3 \text{ and } \gamma = 10,11, \ldots 25\text{ms}\} \tag{3}$$

where $Pq\gamma(r_D)$ is calculated via (2). We use S to derive ROC curves of $Pq\gamma(r_D)$ at different *ISPJ* levels (*q*, γ).

To draw Receiver Operating Characteristics (ROC curves, Figure 3), we need to know the "*true clinical diagnosis*" of the patients. The PSG based diagnosis is considered the absolute truth in this article. The diagnosis based on the test statistic $Pq\gamma(r_D) > P_{th}$ is compared to the "absolute truth," and the nature of the decision is noted as one among (i) true positives (TP), (ii) true negatives (TN), (iii) false positives (FP), or (iv) false negatives (FN). Sensitivity, a measure of success in detecting TPs is defined as: TP/(TP +FN)%.

Specificity, a measure of success in rejecting non-diseased subjects, is defined as TN/(TN +FP)%. The ROC curve is a graph of sensitivity vs. 1-specificity.

The ROC curve (Figure 3) allows us to conclude that snoring carries valuable information on the disease of OSA. The performance of the *ISPJ-probability* indicates the possibility of using the feature to diagnose OSA. Large-database trials are needed for further validation of results.

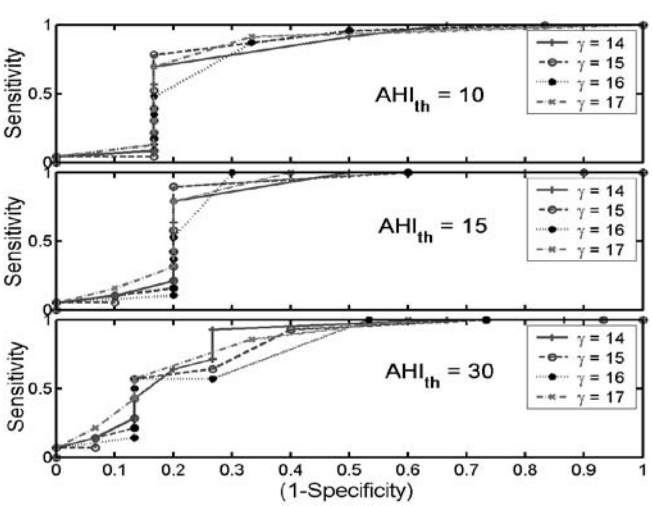
The noncontact nature of the snore acquisition process should provide a significant advantage over competing techniques in developing a community-screening device for OSA. Results obtained in this article have been based on snorers referred to the sleep clinic for suspected OSA. For this reason, the screening utility of the proposed method needs further investigation with healthy snorers who do not display symptoms normally associated with OSA.

The estimation of the TAR, the quantity h(n), can be carried out by using any established technique of blind signal identification (eg., Abeyratne, 1999). This article will not discuss the methodology involved. Interested readers are referred to the article by Karunajeewa (2005).

A Framework for Arousal Analysis in Understanding and Characterizing Sleep Disorders

Electrical activities measured on the scalp, as emanated from the brain, are called EEG signals. EEG is one of the most important signals in overnight PSG recordings. In assessing sleep disorders such as apnea, EEG plays an indispensable role. In the current clinical practice it is being used to study the sleep structure of a patient during sleep, mainly via determining the sleep stages. It is also used to estimate the total sleep time of a patient in computing the AHI and RDI indices.

Figure 3. ROC curves for AHIth =10, 15 and 30 at q=2 and γ = 14−17



In the past, (Deegan & McNicholas, 1995; McNicholas, 1998) researchers have developed various hypotheses and experimental methods to investigate the changes in EEG signal during sleep apnea and hypopnea. Spectral analysis of the EEG has revealed that termination of apnea and hypopnea events are often associated with changes in EEG power in the δ- band of frequency (0.5 – 4Hz) (Dingli & Assimakopoulos, 2002; Svanborg, 1996).

The issue of asymmetry of brain during sleep has been investigated by groups of researchers (Bolduc, 2002; Goldstein & Stoltzfus, 1972; Sekimoto, 2000). Asymmetry studies have been confined to undisturbed sleep in normal individuals. None of these studies considered the significance of apnea/hypopnea on the hemispheric correlations, or how the correlation behaved in the cases of peoples with symptoms of sleep apnea.

Despite the importance of EEG in studying sleep disorders, the main index used to measure the severity of apnea, AHI (or RDI), does not use EEG, except for the total sleep time derived indirectly through EEG. EEG, however, is a high-temporal resolution window to the brain, and contains much more information than what is currently being used in clinical practice. We hypothesize that EEG may provide one avenue to alleviate the important problem mentioned under the Background Section.

Recently, the phenomenon known as the EEG Arousals (EEGA) has received attention as a possible pointer of sleep disorders. EEGA in sleep is defined as an abrupt shift in EEG frequency, lasting for 3s or more. According to the criteria developed by the American Sleep Disorder Association (ASDA), EEGA are markers of sleep disruption and should be treated as detrimental.

Excessive Daytime Sleepiness (EDS) is one of the defining symptoms in OSA patients (Martin, & Wraith, 1997; McNicholas, 1998). The reason for EDS in SDB patients is the modification of sleep texture, due to reoccurring episodes of EEGA causing sleep fragmentation.

In this article, we show that there exists a relationship between the EEGA and information-flow between the left and right hemispheres of the brain, as revealed via the left-right correlation of EEG. This finding may open up a new vista of research in understanding how information flows in the brain during phenomena associated with sleep (eg., EEGA, sleep spindles, and so on) and help with localizing sources of origin of some brain activities. Furthermore, considering the relationship between EEGA and dominant symptoms of OSA (ie., day time sleepiness), we believe a proper understanding of EEG will help in designing a better measure for OSA severity.

Figure 4 shows (Vinayak, 2007) the spectral correlation coefficient computed between EEG signals measured from electrode pairs A1-C4 and A2-C3 (of the International 10/20 electrode system) during an overnight sleep study. In Figure 4, (f) and (g) arousal events and apnea event have been marked to help visualize their relationship with interhemispheric EEG correlation. In Figure 4(a), 4(b), 4(c), and 4(d) correlation coefficients have been computed separately in the δ,θ,α, and β spectral bands of the measured EEG.

Our results showed that apnea/EEG arousal events generally lead to an increase in IHA. This change is most prominent during NREM sleep affected by EEGA, particularly in θ, α, and β frequency bands. Our results strongly suggest that the information flow between the left-right hemispheres of the brain is affected by events of apnea/EEG-arousals (Vinayak, 2007).

FUTURE TRENDS

With the rapidly increasing public awareness of the apnea syndrome, the demand for in-home noncontact methods for screening sleep apnea are expected to rise in the future. Noncontact measurement techniques based on snore (and breathing) sound analysis will, if successful, provide enabling technology for home-monitoring applications.

Once the home-based screening test identifies individuals who should be properly diagnosed, hospital-based PSG testing should be considered. In a hospital-based PSG test, we expect EEG signals to play a pivotal role in the future, in diagnosing sleep apnea. The role of EEG is expected to expand far beyond its current role of sleep-staging (using decades old criteria), making significant inroads into the subject of characterizing apnea. EEG will also help understand the brain mechanisms associated with the processes of sleep and sleep disturbances.

CONCLUSION

In this article, we explored two techniques that addressed a range of difficult problems associated with diagnosing/ screening obstructive sleep apnea. The snore sound-based technique provided a basis for noncontact methods suitable community screening of the disease. Our results indicate that the performance of the feature Intra-Snore-Pitch-Jump Probability (ISPJ-probability) is on par with other competing technologies, in terms of the sensitivity/specificity characteristics. Snore-based diagnosis proposed in this article is superior, in that it does not involve contact instrumentation, thus solving a major problem in population screening of the disease.

We investigated the interhemispheric asymmetry of the brain during apnea and EEG-arousals (EEGA) events. The EEG asymmetry was determined via the Spectral Correlation coefficient of data, which computed the correlation between EEG measured from the two hemispheres of the brain. Results showed that apnea/EEG arousal events generally lead to a decrease in correlation. This change is most prominent during NREM sleep affected by EEGA, particularly in θ, α, and β frequency bands. Our results strongly suggest that the information flow between the left-right hemispheres of the brain is affected by events of apnea/EEG-arousals.

The data used in this article came from: (i) patients medically diagnosed with sleep apnea, and (ii) subjects medically deemed to be without apnea, but nevertheless displaying some symptoms of the disease.

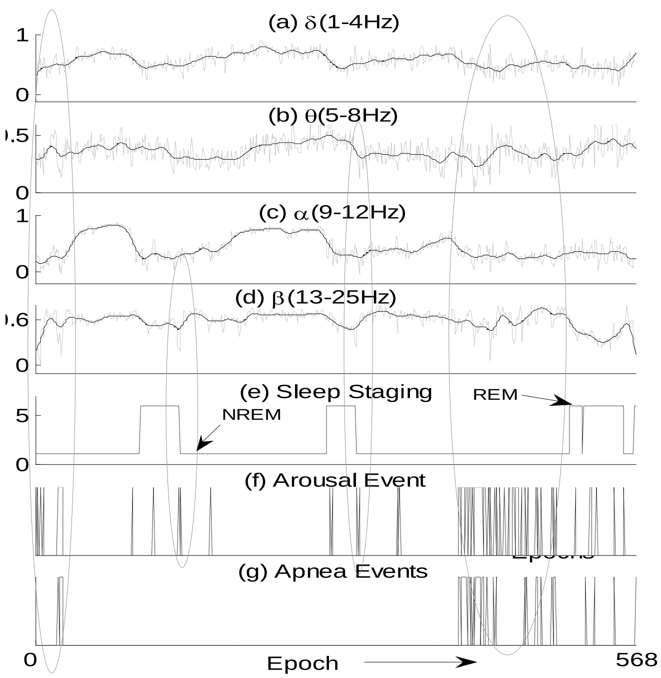
REFERENCES

Abeyratne, U. R., Karunajeewa, A.S., & Hukins, C. (2007). Mixed-phase modelling in snore sound analysis. *Medical & Biological Engineering & Computing, 45*(8), 791-806.

Abeyratne, U. R., Wakwella, A. S., & Hukins, C. (2005). A snore sound based method for the mass screening of sleep apnea. *Physiological Measurement*, *26*, 779–798.

Australian Sleep Association (ASA), Fatigue and Transport Working Party (1999, June). House of Representatives Standing Committee on Communications, Transportation and the Arts. Australia.

Figure 4. (a) Delta, (b) Theta, (c) Alpha, (d) Beta, correlation coefficient during the NREM and REM sleep, (e) dleep stages, base line indicates NREM sleep and 1st level indicates REM sleep, (f) arousal events, (g) apnea events



Babar, S. I., & Quan, S. F. (2002). Through the looking glass, the unrecognized tragedies of obstructive sleep apnea. *Sleep Medicine*, *3*, 299–300.

Bolduc, C., et al. (2002). A study of EEG hemispheric asymmetry during REM sleep in young, healthy adults. *International Journal of Bioelectromagnetism, 2*(4), 145-146.

Deegan, P. C., & McNicholas, W. T. (1995). Athophysiology of obstructive sleep apnoea. *European Respiratory Journal, 8*(7), 1161–1178.

Delaive, R. K., Roos, L., Manfreda, J., & Kryger, M. H. (1998). Obstructive sleep apnea patients use more health care resources ten years prior to diagnosis *Sleep Research Online*, *1*, 71–74.

Dingli, K., & Assimakopoulos, T. (2002). Electroencephalographic spectral analysis: Detection of cortical activity changes in sleep apnoea patients. *European Respiratory Journal 20*(5), 1246–1253.

Ehrlichman, H., Antrobus, J., et al. (1985). EEG asymmetry and sleep mentation during REM and NREM. *Brain and Cognition, 4*, 477–485.

Flemons, J. (2003). Home diagnosis of sleep apnea: A systematic review co-sponsored by the American Academy of sleep medicine. *Chest, 124*(4), 1543–1579.

Flemons, W. W., & Buysse, D. (1999). Sleep-related breathing disorders in adults: Recommendations for syndrome definition and measurement techniques in clinical research. *Sleep*, *22*(5), 667–689.

Goldstein, L., & Stoltzfus, N. (1972). Changes in interhemispheric amplitude relationships in the EEG during sleep. *Physiology & Behavior, 8*, 811–815.

Hoffstein, V. (2000). *Snoring, in the principles and practice of sleep medicine*. In T. Kryger, & W. D. Roth (Eds.), (3rd Ed. pp. 813-826). Philadelphia: W. B. Saunders Co.

Karunajeewa, A. S. (2005). Abeyratne, U. R., & Hukins, C. Mixed-phase modeling of snore sounds within a nonlinear framework for component identification. *IEEE/EURASIP International Workshop on Nonlinear Signal and Image Processing*. Sapporo, Japan.

Kryger, M. H. (2000). Management of obstructive sleep apnea-hyopnea syndrome: Overview. In *The Principles and Practice of Sleep Medicine*. W. B. Saunders Co.

Martin, S., Wraith, P., et al. (1997). The effect of nonvisible sleep fragmentation on daytime function. *American Journal of Respiratory and Critical Care Medicine, 155*(5), 1596–1601.

McNicholas, W. (1998). Arousal in the sleep apnoea syndrome: A mixed blessing? *European Respiratory Journal, 12*(6), 1239–1241.

Portier, F., Portmann, A., Czemichow, L., Vascaut, L., Devin, E., Cuvelier, A., & Muir, J. F. (2000). *American Journal of Respiratory and Critical Care Medicine, 162,* 814–818.

Puvanendran, K., & Goh, K.L. (1999). *Sleep Research Online, 2*(1), 11–14.

Sekimoto, M., et al. (2000). Asymmetric interhemispheric delta waves during all-night sleep in humans. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology, 111*, 924-928.

Svanborg, V., & Guilleminault, C. (1996). EEG frequency changes during sleep apneas. *Sleep, 3*(19), 248-254.

Swarnkar, V., Abeyartne, U., & Hukins, C. (2007). Interhemispheric asynchrony of the brain during events of apnea and EEG arousals. *Physiological Measurement, 28,* 869-880.

Udwadia, Z. F., Doshi, A. V., Lonkar, S. G., & Singh, C. I. (2004). Prevalence of sleep-disordered breathing and sleep apnea in middle-aged urban Indian population. *American Journal of Respiratory and Critical Care Medicine, 169*, 2.

Young, T., Evans, L., Finn, L., & Palta, M. (1997). Estimation of the clinically diagnosed proportion of sleep apnea syndrome in middle-aged men and women. *Sleep, 20*(9), 705–707.

KEY TERMS

β, α, θ and δ EEG Frequency Bands: Bandpass components of EEG signals respectively covering the frequency bands of 18-30 Hz, 8-13Hz, 4-7Hz, and 1-3.5Hz.

EEG Arousals: An abrupt shift in EEG frequencies during sleep which may include α, θ, and/or frequencies <16Hz, excluding structures known as sleep spindles.

International 10/20 Electrode System: A standard placement of electroencephalographic (EEG) electrodes on the scalp to measure signals from the brain.

Oxygen Desaturation: A decrease in the blood oxygen levels often associated with sleep apneas.

Respiratory Disturbance Index (RDI): The number of all respiratory disturbances per hour, averaged over the entire duration of sleep.

Functional Characteristics and Supporting Methods for Maintaining Independent Life of the Elders

Hiromi Nishiguchi
Tokai University, Japan

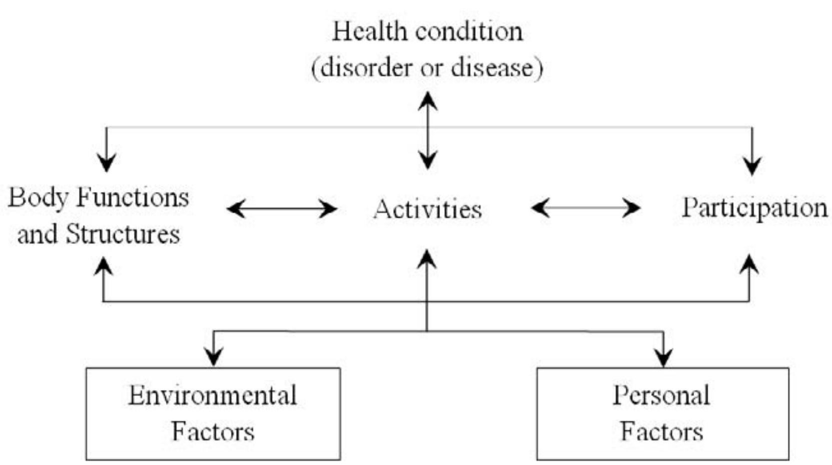
INTRODUCTION

At first, it is necessary to measure physical and intellectual function objectively to support independence of the elders who need care in daily life, and to analyze the reason why the elders cannot move or take action, and why the barrier occurs. An effective use of a survival or substituted function and also the improvement of physical environment, such as housing accommodation in private and public sectors, require removing barriers. In addition, social resources are useful to support independence of the elders in their daily lives. In this chapter, two issues were the focus: functional characteristics that change in accordance with the increase of age, and the methods of removing barriers for elders to move as they like.

BACKGROUND

We take actions with physical and intelligent functions in our daily life, but their functions change as they get old. Some activity limitations or participation limitations will occur when the necessary function for the action declines, as shown in ICF (International Classification of Functioning, Disability and Health) offered by WHO (2001) (cf. Figure 1). Therefore, to support independence of elders who need care in daily life, we must grasp declining status of physical and intellectual functions objectively, and barriers should be removed that cause activity limitations or participation limitations. In addition, an independence of the elders in their community life is feasible by positive utilization of social resources (Nishiguchi, Saito & Ozeki, 1997; 1998).

Figure 1. Interaction between the components of ICF



F

MAIN FOCUS OF THE CHAPTER

Functions of the Elders

How to Grasp Functions of Elders

There are two representative methods to grasp changes of human being functions (cf. Figure 2). One method is the longitudinal survey, which measures functions of one certain elder group as investigation subject. The other is the cross-sectional survey, which is a method to compare functions of different ages.

Change of Physical and Intellectual Functions by Aging

It is generally mentioned that many human being functions change by aging (Charness, 1985). Most physical functions become the highest values at the 20-year-old generation (Nishiguchi, Saito& Ozeki, 1996). On the other hand, it is known that some intellectual function is maintained or raised (Dennis & Tapsfield, 1996). This knowledge was reported by Horn and Cattell (1966), and they classified the intellectual ability into "fluid ability" and "crystallized ability."

Furthermore, Thurstone and Thurstone (1941) proposed that intellectual ability of human beings is composed of "seven ability factors." They are as follows: Memory (M), Number (N), Perception (P), Reasoning (R), Word fluency (W), Space (S), and Verbal comprehension (V).

Cross-Sectional Comparison of Intellectual Ability

The authors made a description-style test to measure the previously mentioned seven ability factors (Nishiguchi & Saito, 2000). This test carried out to two subject groups: one elder group (20 people with an average age of 63.1 years), and a young group (20 people with an average age of 20.9 years).

In this test, four phases of degrees of difficulty (a degree of difficulty rises by order of I→IV) are set for

Figure 2. Cross-sectional and longitudinal survey

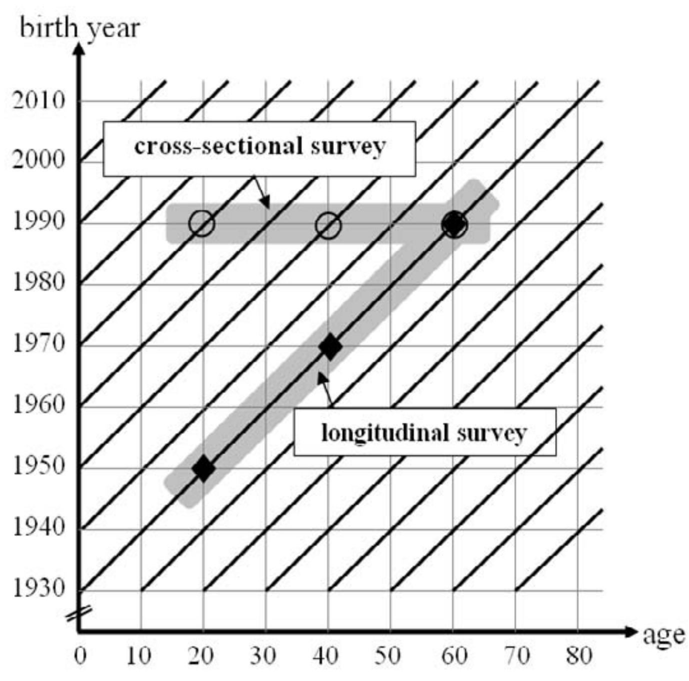
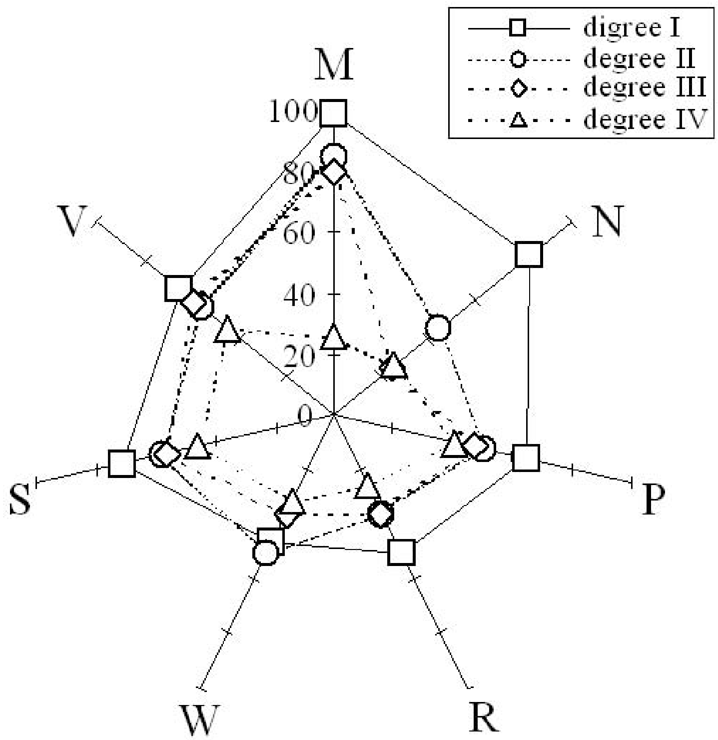


Figure 3. The work performance ratio of the elders group for the youngers



each question. "Work performance" was calculated by the following calculation type.

*work performance(%) = number of correct answers / number of problems * 100*

Figure 3 showed the work performance ratio of the elders group for the young. In two ability factors, ("Memory (M)" and "Number (N)"), the work performance ratio is high, and a difference is not recognized between two groups when a degree of difficulty is low. But the difference between two groups grew bigger as the degrees of difficulty rose. These abilities are "fluid ability," which decline by aging. Furthermore, the elders did not seem to be flexible for a change in difficulty of the tasks comparison with the young.

On the other hand, in Word fluency (W) and Verbal comprehension (V), which are "crystallized abilities," the elders were flexible in changing the difficulty, although they performed lower in comparison with the young.

Activity of Daily Living and Removing Barriers

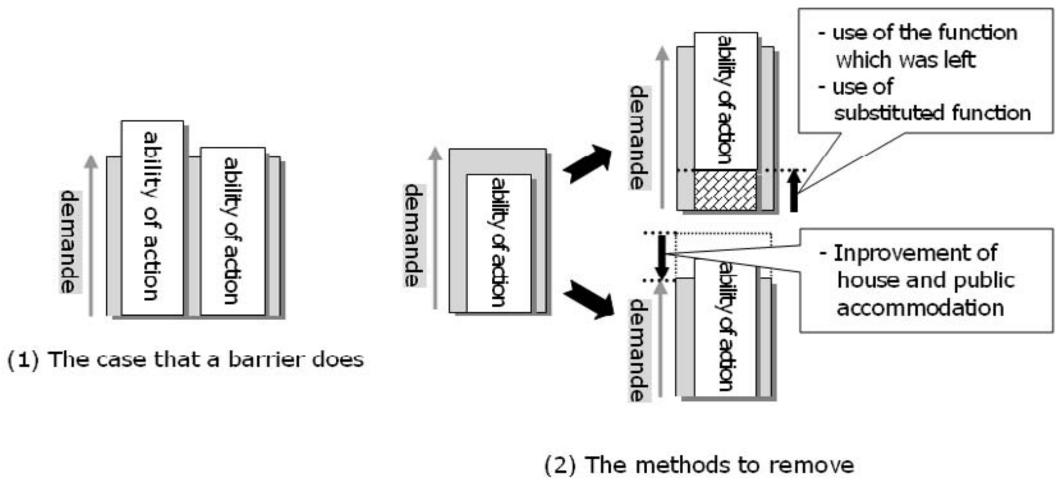
Activity of Daily Living (ADL)

Activity of daily living is the aggregation of movements necessary at the minimum when a human being runs his or her own daily life. ADL consists of seven kinds of body action: lying down or getting up, place movement, eating a meal, excretion, bathing, wearing or taking off clothes, and conditioning a figure. At first, it is necessary to judge whether an elder who needs care can run his or her own daily living to support an independent life. If there is any action that the elders cannot do (i.e., "barrier"), a discussion is necessary to remove their barriers and support them.

Methods to Remove Barriers

In the situation that someone is going to accomplish some actions in everyday life, a barrier occurs when his or her ability does not satisfy the ability required, as shown on the right side as (2) of Figure 4. When a

Figure 4. Methods to remove barriers



survival or substituted function can be used, the barrier will be removed by using some jig or self-help device (Nishiguchi, Saito & Ozeki, 1993), and then, the objective action becomes easy to perform for the elders who need care. In addition, it is effective to apply an architectural barrier-free design to housing and public accommodation.

FUTURE TREND

Institution care or home care is usually provided for the elders who need care in daily life. In that case, it is very effective to utilize "social resources" such as a welfare service providing system, welfare manpower (e.g., a social worker, care worker, local volunteer), welfare technical support (e.g., welfare machinery/tool). Providing higher quality social resources by developing health care networking systems are expected in the future.

CONCLUSION

In this chapter, focus was placed on the changes in functional characteristics by aging and methods of removing barriers in elders' daily lives. Functions of human beings change by aging, but ways of change are different in the two functions. Many physical functions become the highest values at the 20-year-old generation, but some intellectual functions, "crystallized abilities," is maintained, or rather raised. Particularly, when a physical function deteriorates, a daily living activity will become difficult, and barriers will occur. In removing a barrier, it is indispensable to use survival or substituted function and improve living environment. It is of importance that good social resources are provided to the elders and they are able to live independently, based on their own needs and faith in the future (Rouse, 1993).

REFERENCES

Charness, N. (1985). *Aging and human performance*. John Willy & Sons. Inc.

Dennis, I., & Tapsfield, P. (1996). *Human abilities—Their nature and measurement*. Lawrence Erlbaum Associates, Publishers.

Horn, J.L., & Cattell, R.B. (1966). Refinement and test of theory of fluid and crystallized intelligence. *Journal of Educational Psychology, 57*, 253–270.

Nishiguchi, H., Saito, M., & Ozeki, M. (1993). A study on the evaluation of functionality and effectiveness of self-help devices for the disabled. *Journal of Japan Industrial Management Association, 44*(3), 200–208.

Nishiguchi, H., Saito, M., & Ozeki, M. (1996). The ability of the aged to detect errors in the intellectual work—The influences of the personality. *The Japanese Journal of Ergonomics, 32*(2), 99–104.

Nishiguchi, H., Saito, M., & Ozeki, M. (1997). A study of the recreational activity in the aged generation—The influence of the habits and the personality. In *Proceeding of the 13th Triennial Congress of the International Ergonomics Association*.

Nishiguchi, H., Saito, M., & Ozeki, M. (1998). A study of the recreational activity in the aged generation—The influence of the physical condition and the habits. In *Proceedings of the 5th International Congress of Behavioral Medicine.*

Nishiguchi, H., & Saito, M. (2000). A study of the intellectual ability of the aged—The relationship between the degree of difficulty and performance. In *Proceedings of the 14th Triennial Congress of the International Ergonomics Association, 99*.

Rouse, W.R. (1993). *Catalyst for change, concepts and principles for enabling innovation*. John Wiley & Sons, Inc.

Thurstone, L.L., & Thurstone, T.G. (1941). *Factorial studies of intelligence*. University of Chicago Press.

World Health Organization. (2001). International classification of functioning, disability and health.

KEY TERMS

Activity of Daily Living (ADL). ADL consists of seven kinds of body actions, and there is IADL (Instrumental ADL), which is another classification of activities in daily life. IADL is composed of activities requiring intellectual ability, such as money management, telephone calls, housework (shopping, meal preparation, cleaning).

ICF: ICF was defined by WHO (World Health Organization) in 2001. This classification is the revised version of ICIDH (International Classification of Impairments, Disabilities and Handicaps), which was composed from a viewpoint of disorder. In the ICF model, human behavior was classified into three categories (body functions and structures, activities, and participation).

Independence in Daily Life. Independence in daily life means the condition in which a person with a disorder can live according to his or her own will (influenced by needs and belief).

Seven (Ability) Factors: Details of "seven factors" proposed by Thurstone and Thurstone (1941) are as follows: Memory—the ability to memorize and recall; Number—the ability to solve arithmetic problems; Perception—the ability to see differences and similarities among objects; Reasoning—the ability to find rules; Word fluency—the ability to produce words rapidly; Space—the ability to visualize relationships; and Verbal comprehension—the ability to define and understand words.

Social Resources: In welfare services, social resources are indispensable and have to be easy to use for elders who need care or for their families.

Survival and Substituted Function: A survival function is a function that has been left a little after disease, and it can be used effectively with some jig or self-help device. A substituted function has a supplementary roll for the function with disorder, such as an auditory function in the blind.

Generating Simulated DT-MRI Dataset

Dilek Goksel Duru
Bogazici University, Turkey

Mehmed Ozkan
Bogazici University, Turkey

INTRODUCTION: DIFFUSION TENSOR IMAGING

Diffusion tensor magnetic resonance imaging (DT-MRI), also known as diffusion tensor imaging (DTI), needs post processing by adequate image analysis and visualization tools. White matter tractography using DTI is becoming a routine MR technique to study white matter properties, connectivity, and alterations of fiber integrity due to pathology. The success of the method depends on the accuracy of the tracking algorithms. The disadvantage on the evaluation of such methods is that there is no gold standard regarding the true geometry of the brain anatomy or fiber bundles reconstructed in each particular case.

Considering the increasing number of reports on DTI post-processing research, it is observed that most of the methodology is applied only to real data, acquired from human or animal test subjects. This approach has the following drawbacks: (i) the true anatomy of each imaged subject is not known in detail, (ii) the image noise is of little control, (iii) cost of scanner time is usually high, (iv) the comparison of the methods and the results from different studies is difficult because the data has been acquired on various MR scanner hardware, and often scanned with different imaging protocols, and finally, (v) the test data is collected from different subjects, each having unique characteristics in gross anatomy and tissue micro-architecture. In order to reduce these problems, a synthetic DTI dataset with known geometric and signal properties has been developed.

The accuracy of white matter anatomical maps obtained by DTI is still unclear due to the general inability of the diffusion tensor model describing a single voxel with multiple orientational maxima. Verification and validation of the synthetic data analysis' aims to elucidate the white matter fiber tractography in eliminating the uncertainty areas and understanding the connectivity more clearly and reliable.

A model must describe how water diffuses in the synthetic dataset. For simplicity, we have in this work considered only two very basic models. In this work two separate models with different geometric properties characterized by anisotropic Gaussian diffusion are specified. These models are sampled and their output is similar to those obtained from MR scanners are generated. The efficient calculation of the diffusion tensor is achieved from that output. It is used to generate several common measures and visualizations describing Gaussian water diffusion. The project covers the geometric model, discrimination, sampling, tensor calculation, parameter calculation, and the visualization.

BACKGROUND

DT-MRI Pulse Sequences: Encoding for Diffusion

Diffusion-weighted images are the raw data source used to calculate the diffusion tensor. In DTI, each voxel is assigned a tensor that describes local water diffusion. The relationship between the loss of phase coherence in the transverse spin radio frequency (RF) signal S_0 and S_i, and the gradient pulse $g=[g_{ix}\, g_{iy}\, g_{iz}]^T$ with $\hat{g}_i^T D \hat{g}_i$ the apparent diffusivity along g_i is given by the Stejskal-Tanner equation (Stejskal, 1965):

$$S_i = S_0 e^{-b\hat{g}_i^T D \hat{g}_i} \qquad (1)$$

A Stejskal-Tanner imaging sequence may be implemented by adding diffusion encoding gradients to standard anatomical MRI pulse sequences (Ciccarelli et al., 2003).

By systematically applying diffusion gradients in multiple directions, a mathematical construct known as the diffusion tensor, D, could be estimated at each point in the tissue. The utility of the diffusion tensor is that it provides the direction in three dimensional space in which the rate of diffusion is greatest (Basser, Pajevic, Pierpaoli, Duda, & Al-droubi, 2000; Borisenko & Tarapov, 1979).

Estimation of the Diffusion Tensor

Derivation of structural information follows a measured displacement characteristic related by means of a model to the physical and geometrical properties of the tissue. Diffusion coefficients and shapes of semipermeable membranes of compartments in the system are these related characteristics. The behavior of the MR signal and the measured *apparent diffusion coefficient* (ADC) as anisotropic diffusion indexes are greatly affected by the cellular architecture of a tissue, mainly because cellular membranes are relatively impermeable to water.

The relationship between loss of phase coherence in the transverse spin RF signal and the gradient pulse *g* is given by the Stejskal-Tanner equation (1), where *b* is the diffusion weighting factor (Ciccarelli et al., 2003) given by:

$$b = \gamma^2 \delta^2 \left[\Delta - \left(\frac{\delta}{3}\right) \right] |g|^2 \qquad (2)$$

Here, γ is the gyromagnetic ratio, δ is the gradient pulse width, Δ is the time between the gradient pulses, $|g|$ is the strength of the diffusion gradient pulses.

Basser (Basser, Mattiello, & Le Bihan, 1994), building on the work of Stejskal and Tanner (Stejskal, 1965), has shown that the diffusion tensor can be calculated from knowledge of signal attenuation and magnetic gradient strengths applied during a diffusion weighted spin echo experiment using the following equations;

$$\ln\left(\frac{A(b)}{A(b=0)}\right) = -\sum_{i=1}^{3}\sum_{j=1}^{3} b_{ij} D_{ij} = -\left(b_{xx}D_{xx} + 2b_{xy}D_{xy} + 2b_{xz}D_{xz} + b_{yy}D_{yy} + 2b_{yz}D_{yz} + b_{zz}D_{zz}\right) \qquad (3)$$

$$Trace(D) = D_{xx} + D_{yy} + D_{zz} = 3\langle D \rangle = \lambda_{xx'} + \lambda_{yy'} + \lambda_{zz'} \qquad (4)$$

where *A(b)* is the voxel attenuated signal (echo) intensity recorded in the presence of gradients (3), *A(0)* is the gradient-free, unattenuated echo intensity, D_{ij} is the (symmetric, positive definite, 3 by 3) diffusion tensor (3), and b_{ij} is a matrix specified by the magnetic field gradients applied during the spin echo. In eq. 3:

$$\sum_{i=1}^{3}\sum_{j=1}^{3} b_{ij} D_{ij} \equiv b : D$$

is the standard scalar product of two tensors. This so called *b-matrix* (2, 3) has the form: $A = A_0 e^{(-b:D)}$.

Tensor Analysis and the Diffusion Tensor: PCA

Principal component analysis (PCA) is a classical statistical method widely used in data analysis and compression. PCA is based on the statistical representation of a random variable. The method reduces data dimensionality by performing a covariance analysis between factors. PCA method is based on linear transformations; however, nonlinear extensions exist. PCA is a technique for reducing second-order dependencies in the data by rotating the axes to correspond to orthogonal directions of maximum covariance (decorrelation).

From a symmetric matrix such as the covariance matrix, an orthogonal basis by finding its eigenvalues and eigenvectors can be calculated. The diffusion tensor *D* is a real, symmetric second order tensor, represented in matrix form as a real, symmetric 3x3 matrix (5).

$$D = \begin{bmatrix} D_{xx} & D_{xy} & D_{xz} \\ D_{yx} & D_{yy} & D_{yz} \\ D_{zx} & D_{zy} & D_{zz} \end{bmatrix} \qquad (5)$$

The diagonalization of the positive definite and symmetric diffusion tensor results in a set of three eigenvalues, $\lambda_1, \lambda_2, \lambda_3$, listed in decreasing order. The eigenvectors e_i and the corresponding eigenvalues λ_i are the solutions of the diagonalization of *D* (5), where the eigenvectors e_i are the principal diffusion directions $e_i (i = 1, 2, 3)$. The eigensystem of the diffusion tensor may be interpreted graphically as an ellipsoidal surface with semimajor axis oriented in the e_1 direction and semiminor axis oriented in the e_2 and e_3 directions regarding to $De_i = \lambda_i e_i$.

The lengths of the axis in this ellipsoidal interpretation are given by the corresponding eigenvalues of each eigenvector, with semimajor axis length λ_1 and semiminor axis lengths λ_2 and λ_3. In order for the analogy between the symmetric real tensor and the ellipsoid to be physically realizable, the eigenvalues of *D* must be nonnegative (Bammer, 2000; Pajevic & Pierpaoli, 1999). In cases of purely isotropic diffusion, the diffusion ellipsoid takes on a spherical shape, as $\lambda_1 = \lambda_2 = \lambda_3$. The two extreme cases of physically realizable anisotropic diffusion where the ellipsoidal description degenerates are purely linear anisotropic diffusion, $\lambda_1 = c$, and $\lambda_2 = \lambda_3 = 0$, where the diffusion ellipsoid degenerates into a line pointing in the e_1 direction, and the purely planar anisotropic diffusion, where the diffusion ellipsoid becomes oblate, meaning that $\lambda_1 = \lambda_2$, $\lambda_3 = 0$.

Using the three-dimensional Gaussian Stejskal-Tanner model, the six unique elements of the diffusion tensor *D* may be solved by acquiring at least six diffusion weighted measurements in noncollinear measurement directions *g* along with a nondiffusion-weighted measurement S_0 (6). Taking more than six diffusion weighted measurements creates an over constrained system of equations which may be solved using least square methods (Basser, 2002; Basser et al., 2000). The advantage of over-constraining the solution for *D* is a reduction in the amount of noise propagating from diffusion weighted measurements S_i into the calculated diffusion tensor. The linear system of $n \geq 6$ diffusion weighted measurements constraining the diffusion tensor *D* may be represented in matrix form as in Equation 6.

$$\begin{bmatrix} x_1^2 & y_1^2 & z_1^2 & 2x_1y_1 & 2y_1z_1 & 2x_1z_1 \\ x_2^2 & y_2^2 & z_2^2 & 2x_1y_1 & 2y_2z_2 & 2x_2z_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_n^2 & y_n^2 & z_n^2 & 2x_ny_n & 2y_nz_n & 2x_nz_n \end{bmatrix} \begin{bmatrix} D_{xx} \\ D_{yy} \\ D_{zz} \\ D_{xy} \\ D_{xz} \\ D_{yz} \end{bmatrix} = \begin{bmatrix} -\frac{1}{b}\ln\frac{S_1}{S_0} \\ -\frac{1}{b}\ln\frac{S_2}{S_0} \\ \vdots \\ -\frac{1}{b}\ln\frac{S_n}{S_0} \end{bmatrix} \quad (6)$$

In the linear system of equations $Ad = s$ (6), A is the encoding matrix containing the $n \geq 6$ unit normalized gradient measurement directions, *d* is a vector specifying the 6 unique elements of the diffusion tensor *D*, and s is a vector containing natural logarithmic scaled RF signal loss resulting from the Brownian motion of spins.

Although not commonly mentioned in DTI literature, the existence of nonphysically realizable diffusion tensors must sometimes be accounted for as a special case in tractography algorithms (Basser et al., 2000; Hasan & Narayana, 2003; Poupon et al., 2000; Taylor, 2004; Westin, Maier, Khidhir, Everett, Jolesz, & Kikinis, 1999; Taylor & Hsu, 2004).

Creating the Synthetic DT-MR Dataset

In order to study, validate, and compare various denoising and fiber tracking methods, a mathematical phantom consisting of various tract geometries with well-known diffusion characteristics is needed. A model must, for each voxel, provide an orthogonal set of directions of water diffusion as well as a measure of the amount of diffusion in each of these directions.

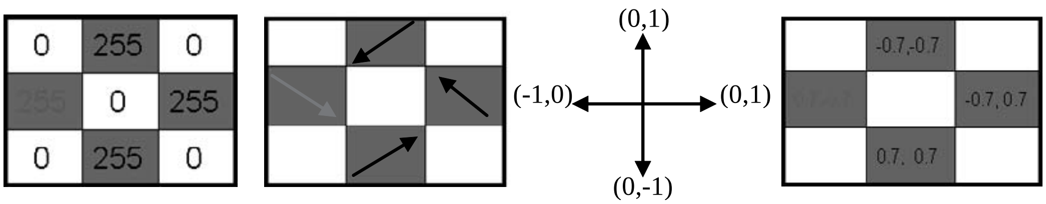
The diffusion tensor is a symmetric positive semidefinite matrix, at least in the absence of noise. Therefore it is known that the eigenvectors of *D* form an orthogonal basis for the diffusion in each voxel and that the eigenvalues are positive. The diffusion is dominant in the direction corresponding to the largest eigenvalue, meaning *principal diffusion direction*. Where the analysis is based on the PCA as explained in a previous chapter.

One common way of visualizing both the eigenvectors and the eigenvalues is to let the eigenvectors be the basis of an ellipsoid where the distance from the center to the surface in each basis direction is equal to the eigenvalue of the corresponding eigenvector.

Gram Schmidt Orthogonalization

Assuming that the b factor in the matrix form b=[bxx byy bzz bxy bxz byz] is [1 0 0 0 0 0], it is obvious that the diffusion tensor (*D*) will have only one element with the value D_{xx}. *D* should be calculated in order to generate the image data *S* for each pixel *(i, j)*. As explained previously, *D* consists of the eigenvectors and eigenvalues of a given pixel *(i,j)*. $D = U\Sigma V$, where *U* contains the eigenvectors, and Σ contains the eigenvalues. As an example, taking the pixel (2, 1) into account, assume the eigenvalue is [6, 0, 0], the eigenvector for pixel (2, 1) is, [0.707, –0.707, 0], so *U* can be found by Gram-Schmidt orthogonalization.

Figure 1. The pixelwise representation of the calculation explained in Equation (8)



Creating $D = [Dxx, Dyy, Dzz, Dxy, Dxz, Dyz] = [2.9, 2.9, 0, –2.9, 0, 0]$, the image data will be calculated as in (7) regarding to (1):

$$S(i, j) = T_2(i, j)e^{-bD} \quad (7)$$

$$S(2,1)=T_2(2,1) * \exp(-b.D) = 255 * \exp(-2.9)= 255 * 0.05 = 14.03 \quad (8)$$

Given a set of k linearly independent vectors $\{\underline{v}_1, \underline{v}_2, \ldots \underline{v}_k\}$ that span a vector subspace V of R^n, the Gram-Schmidt process generates a set of k orthogonal vectors $\{\underline{q}_1, \underline{q}_2, \ldots, \underline{q}_k\}$ that are a basis for V.

$$\hat{q}_i^T \hat{q}_j = \begin{cases} 1 & \text{for } i = j \text{ (unit norm)} \\ 0 & \text{for } i \neq j \text{ (orthogonality)} \end{cases}$$

$D = U \Sigma V^t =$

=

0.7	0.7	0
-0.7	0.7	0
0	0	1

6	0	0
0	0	0
0	0	0

0.7	-0.7	0
0.7	0.7	0
0	0	1

=

2.9	-2.9	0
-2.9	2.9	0
0	0	0

RESULTS

After post processing the donut-shaped synthetic data, the original values and the calculated results do match, so the algorithm determines the diffusion process and the tracking path successfully (Figure 2). Also the synthetic data is examined for different signal to noise ratio (SNR) values, the results for SNR=10 are shown Figure 3. The developed toolbox analyzing different types of geometries like kissing, branching, and linear trajectories with different SNR values is validated. The tractography algorithm may be followed after that verification, and a validated fiber tractography may be visualized.

The b factor in our study, which gives information about the direction and amplitude of the diffusion gradient, is a 1x7 matrix in form $B = [B0, Bxx, Byy, Bzz, Bxy, Bxz, Byz]$. In the self created synthetic data, 6 diffusion weighted images and a null image namely the T2 image creating a set of intensity images of size 256x256x7 is generated for the analysis, and in the real human brain dataset (Figure 4), the matrix is of size 256x256x31. In a human myocardium sample, 10 diffusion weighted images and the T2 image creating a set of intensity images of size 256x256x11 is used for the analysis (Figure 5). The relation can be expressed generally; $B(n,:)$ has the information of the intensity image $S(:,:,n)$. Based on the Stejskal Tanner equation, the diffusion tensor $D = [Dxx, Dyy, Dzz, Dxy, Dxz, Dyz]$ is calculated in the algorithm. For the validation of the algorithm, real data sets of human cardiac and brain diffusion MR images are used.

FUTURE TRENDS

In a future work, the fiber tractography will be accomplished via a bootstrap method for the estimation of the dispersion associated with pathological tissue, where the data are drawn with replacement; many more feasible DT-MRI volume realizations can be extracted than with a one acquisition one data set approach. With this approach, likelihood maps of the tracts from the pathological brain region of interests will be created. The algorithm will also be used as a probabilistic tractography method to determine connection likelihoods between two or more DT regions of interest. The

Figure 2. The donut shaped synthetic data: Blue (outer border) represents original values, red (filled in) calculated results; they are totally overlapping

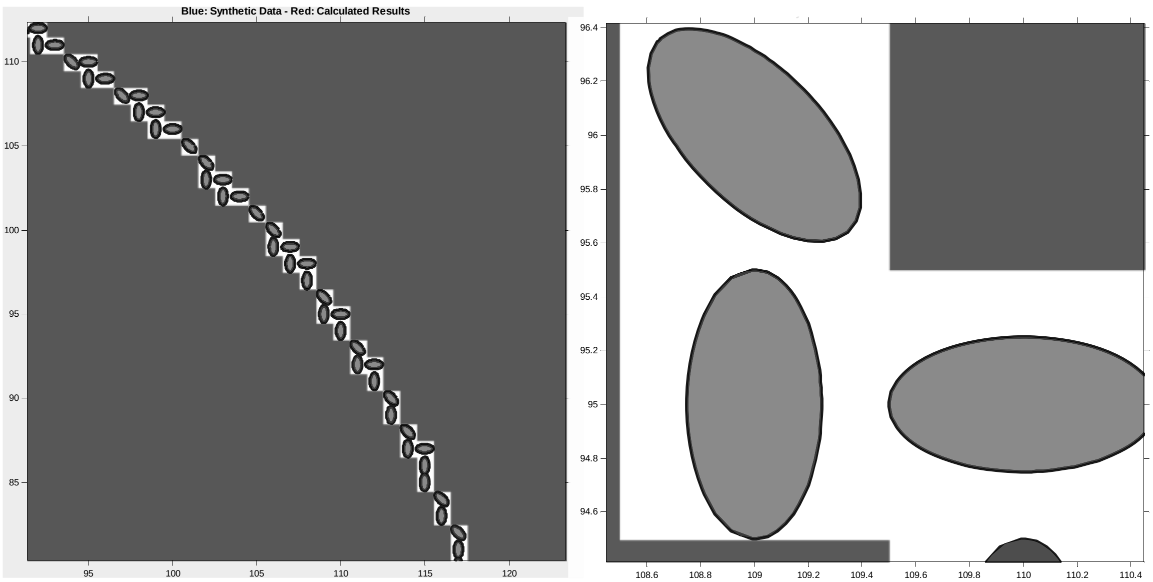


Figure 3. The donut shaped synthetic data (SNR=10)

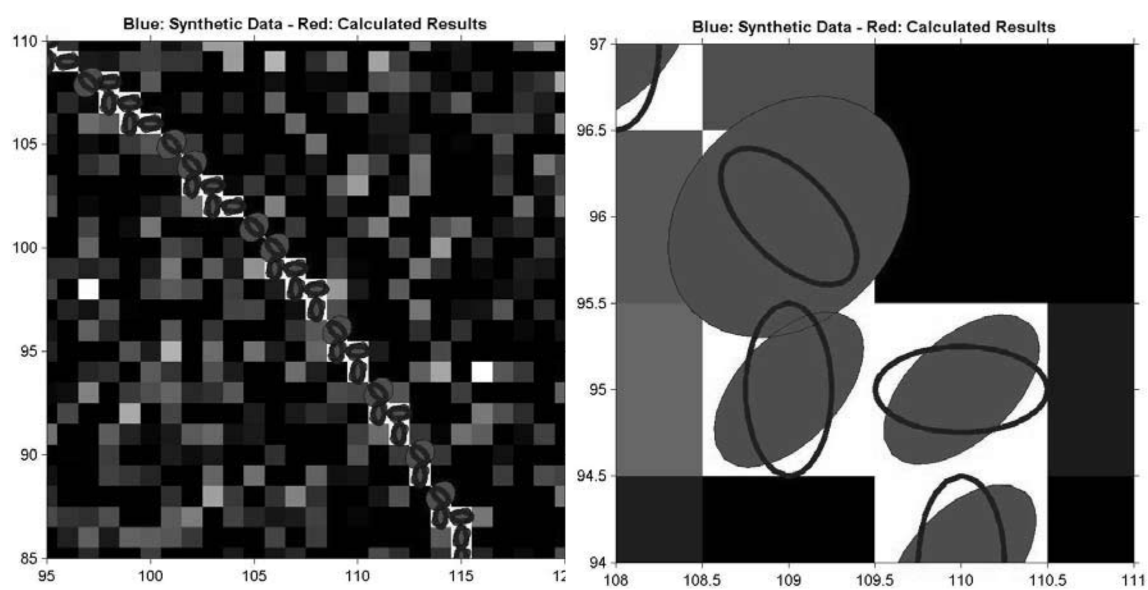


Figure 4. Real data. Human Brain: Principal eigenvector of the diffusion tensor overlaid on ADC map. Outside the brain a chaotic distribution of the tensor data is determined, whereas in the brain region the diffusion tensor is properly distributed.

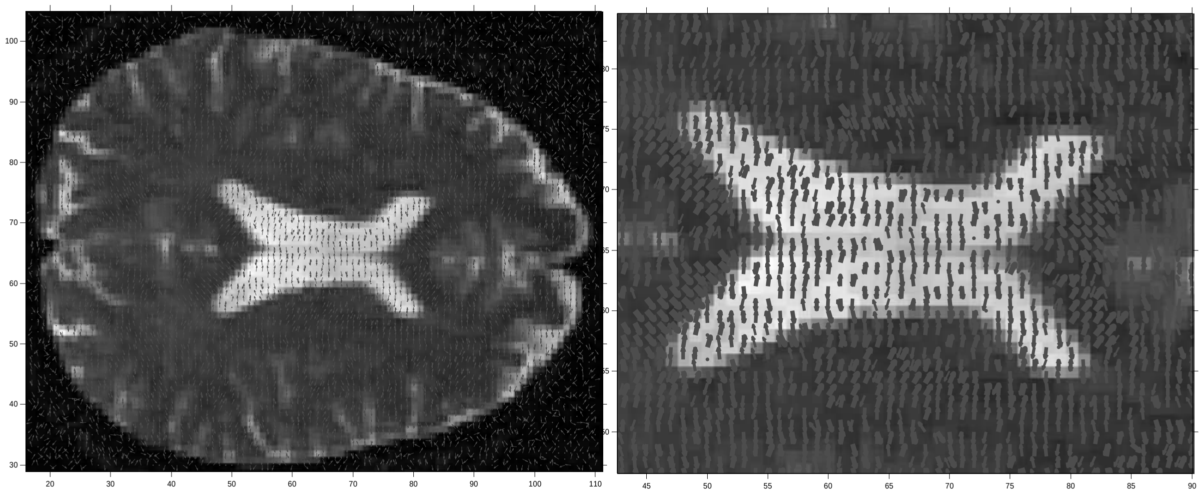
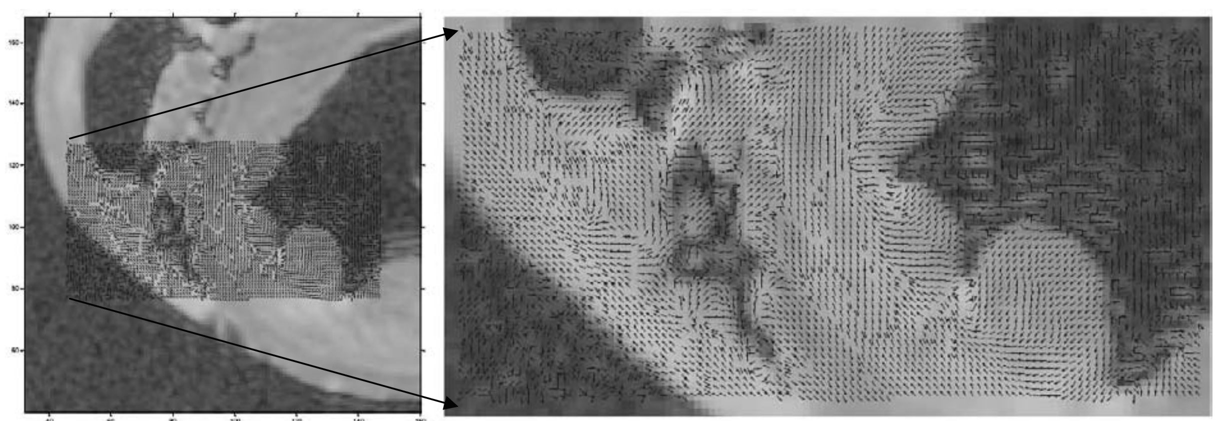


Figure 5. Real data. Human Myocardium: Principal eigenvector of the diffusion tensor overlaid on ADC map of the myocardium. Region of interest is zoomed on the right, proper distribution of the principal eigenvectors of the diffusion tensor is determined.



tracking problems in the uncertainty areas as crossing, branching, and so forth are aimed to be solved regarding the upgrading of the proposed method where the calculation will be implemented on a neural system basis and physiological background to allow better understanding of how impairment in neural circuitry is associated with neuropsychiatric disease.

CONCLUSION

There are still many drawbacks in the literature on the analysis of DTI, especially in neurological disorders. The classification of brain white matter fiber tracking is of great importance. One of the important goals is to better describe the apparent diffusion process in these multiple fiber regions, thus overcoming the limitations of classical DTI. Developing a reliable and rapid tractography tool for the clinical use based on probabilistic approach determines the major part of the future study of the work in progress in that sense. Diffusion tensor imaging is limited in its ability to accurately describe local tract orientation in cases of branching or crossing structure. Because the ellipsoidal geometry associated with the Gaussian tensor model has only one orientational maximum, the tensor model is inadequate when multiple tract orientations exist within a single voxel (Bammer, 2003; Basser et al., 1994; Barisenko, 1979; Basser et al., 2000; Berg, 1983; Ding, Gore, & Anderson, 2003; Ulug & van Zijl, 1999). Mathematical formalisms for describing the information content of the diffusion tensor have been developed by different groups (Ciccarelli et al., 2003; Hasan & Narayana, 2003; Lori, Akbudak, Shimony et al., 2002). The quality of the estimated parameters depends strongly on the strengths and directions of diffusion-encoding gradients (Ding et al., 2003) and the noise efficiency of different gradient configurations (Kindlmann, Weinstein, & Hart, 2000; Mangin, Poupon, Clark, Le Bihan, & Bloch, 2003; Masutani, Aoki, Abe, Hayashi, & Otomo, 2003; Ozarslan, Vemuri, & Mareci, 2004).

In our study, first the diffusion tensor MR data is simulated, and this is followed by the diffusion tensor analysis. In the proposed study, the algorithm is verified on phantom DT images. The diffusion tensor analyzing toolbox is written in Matlab. The PCA helps to get the eigenvectors and eigenvalues of the system, and these denote the diffusion tensor. In our experimental data, circular, kissing, branching, crossing, linear, and curved trajectories with different SNR values are analyzed, and the results of the synthetic data and the known values do match, as seen in Figure 2 and Figure 3, where blue represents the original, known values and red the calculated results. Regarding this validation and verification, the tool is applied to real human data (Figure 4 and Figure 5). To conclude, the simulated data experiment on diffusion tensor analysis has succeeded. As a further conclusion, according to these validated results, a verified tractography may be done, which is still missing in the literature.

ACKNOWLEDGMENT

The authors thank P. Helm and Stefan Skare for the real DT-MRI data.

REFERENCES

Bammer, R. (2000). *Fast and ultrafast magnetic resonance diffusion-weighted imaging in the presence of physiologic motion.* Submitted to the Faculty of Electrical Engineering in fulfillment of the requirements for the degree of Doctor of Technical Sciences at the Graz University of Technology.

Bammer, R. (2003). Basic principles of diffusion-weighted imaging. *European Journal of Radiology, 45,* 169-184.

Basser, P.J. (2002). Relationships between diffusion tensor and q-space MRI. *Magnetic Resonance in Medicine, 47,* 392-397.

Basser, P.J., Mattiello, J., & Le Bihan, D. (1994). MR diffusion tensor spectroscopy and imaging. *Biophys. J., 66,* 259-267.

Basser, P.J., Pajevic, S., Pierpaoli, C., Duda, J., & Al-droubi, A. (2000). In vivo fiber tractography using DT-MRI data. *Magnetic Resonance in Medicine, 44,* 625-632.

Berg, H.C. (1983). *Random walks in biology.* Princeton, NJ: Princeton University Press.

Borisenko, A.I., & Tarapov, I.E. (1979). *Vector and tensor analysis with applications*. New York: Dover Publications, Inc.

Ciccarelli, O., Parker, G.J.M., Wheeler-Kingshott, C.A.M., Barker, G.J., Boulby, P.A., Miller, D.H., et al. (2003). From diffusion tractography to quantitative white matter tract measures: A reproducibility study. *NeuroImage, 18,* 348-359.

Ding, Z., Gore, J.C., & Anderson, A.W. (2003). Classification and quantification of neuronal fiber pathways using diffusion tensor MRI. *Magnetic Resonance in Medicine, 49*, 716-721.

Hasan, K.M., & Narayana, P.A. (2003). Computation of the fractional anisotropy and mean diffusivity maps without tensor decoding and diagonalization: Theoretical analysis and validation. *Magnetic Resonance in Medicine, 50,* 589-598.

Kindlmann, G., Weinstein, D., & Hart, D. (2000). Strategies for direct volume rendering of diffusion tensor fields. *IEEE Transactions on Visualization and Computer Graphics, 6,* 124-138.

Lori, N.F., Akbudak, E., Shimony, J.S., et al. (2002). Diffusion tensor fiber tracking of human brain connectivity: Acquisition methods, reliability analysis and biological results. *NMR Biomed., 15,* 493-515.

Mangin, J.F., Poupon, C., Clark, C., Le Bihan, D., & Bloch, I. (2003). Distortion correction and robust tensor estimation for MR diffusion imaging. *Medical Image Analysis, 6,* 191-198.

Masutani, Y., Aoki, S., Abe, O., Hayashi, N., & Otomo, K. (2003). MR diffusion tensor imaging: Recent advance and new techniques for diffusion tensor visualization. *European Journal of Radiology, 46,* 53-66.

Ozarslan, E., Vemuri, B.C., & Mareci, T.H. (2004). Fiber orientation mapping using generalized diffusion tensor imaging. In *ISBI 2004 From Nano to Micro* (Arlington).

Pajevic, S., & Pierpaoli, C. (1999). Color schemes to represent the orientation of anisotropic tissues from diffusion tensor data: Application to white matter fiber tract mapping in the human brain. *Magnetic Resonance in Medicine, 42*, 526-540.

Poupon, C., Clark, C.A., Frouin, V., Regis, J., Bloch, I., Le Bihan, D., et al. (2000). Regularization of diffusion-based direction maps for the tracking of brain white matter fascicles. *NeuroImage, 12,* 184-195.

Stejskal, E.O. (1965). Use of spin echoes in a pulsed magnetic-field gradient to study anisotropic, restricted diffusion and flow. *J. Chem. Phys., 43*, 3597-3603.

Taylor, A.J. (2004). *Diffusion tensor imaging: Evaluation of tractography algorithm performance using ground truth phantoms*. Thesis Submitted to the Graduate Faculty of the Virginia Polytechnic Institute and State University.

Taylor, W.D., Hsu, E., Krishnan, R.R., & MacFall, J.R. (2004). Diffusion tensor imaging: Background, potential, and utility in psychiatric research. *Biological Psychiatry, 55,* 201-207.

Ulug, A.M., & van Zijl, P.C.M. (1999). Orientation-independent diffusion imaging without tensor diagonalization: Anisotropy definitions based on physical attributes of the diffusion ellipsoid. *Journal of Magnetic Resonance Imaging, 9,* 804-813.

Westin, C.F., Maier, S.E., Khidhir, B., Everett, P., Jolesz, F.A., & Kikinis, R. (1999). *Image processing for diffusion tensor magnetic resonance imaging*. Medical Image and Computer-Assisted Intervention MICCAI '99, Lecture Notes in Computer Science (pp. 441-452).

KEY TERMS

Anisotropic Diffusion: Any restriction in the Brownian movements of the free water protons due to physiological obstacles or any other anatomical barriers in a tissue.

Brownian Motion: The random movement of particles suspended in a fluid and the mathematical model used to describe such random movements.

Diffusion Coefficients: Characteristic parameters of the diffusion process such as apparent diffusion coeffient, fractional and relative anisotropy, linear, planar, and spherical anisotropy.

Fiber Tractography: The process of using DTI data to estimate white matter structures.

Gram-Schmidt Orthogonalization: A procedure which takes a nonorthogonal set of linearly independent vectors and constructs an orthogonal vector basis.

Principal Diffusivity: The dominant diffusion direction determined by the largest eigenvector of the diffusion tensor.

Spin Echo (SE) Imaging Sequence: MRI imaging pulse sequence generated by applying 90^o and 180^o pulses for dephasing and rephasing the spins before the readout.

Stejskal Tanner Imaging Sequence: Spin-echo MRI imaging sequence with the addition of two diffusion-encoding gradients right before and after the 180^o RF pulse.

Guided Interactive Diagnostic Assistance

Giovanni Maria Sacco
Università di Torino, Italy

INTRODUCTION

One of the significant challenges and opportunities in computer-assisted healthcare is how to improve diagnosis and treatment of patients. With respect to the past, diagnosis and treatment have become more and more difficult because of two major reasons: a global society where people, goods, and, consequently, diseases travel very easily; and a growing awareness and emergence of very rare "orphan" pathologies that affect an extremely small percentage of the population. Physicians are very often required to diagnose and correctly treat diseases that they are not familiar with. Especially in this context, but also with more frequent pathologies, computer assistance can become an essential tool in healthcare, especially if it supports all relevant information such as protocols, exceptions, best practices, condition-specific guidelines, and so on, in addition to diagnosis.

BACKGROUND

Medical diagnostic systems have a relatively long history with prototype systems developed as early as the '70s. The emphasis was on Artificial Intelligence (AI) techniques with an evolution from systems in which diagnostic knowledge from experts was captured in the form of empirical classification rules (Buchanan & Shortliffe, 1984) to theoretical and model-based approaches, such as set-covering (Reggia, Nau, & Wang, 1983), abductive diagnosis (Console, Theseider Dupré, & Torasso, 1989), and consistency-based diagnosis (de Kleer, Mackworth, & Reiter, 1992).

Regardless of the theory of diagnosis used, a dramatic limitation of most diagnostic systems is their system-centric, rather than user-centric approach. This implies a master-slave relationship in which the system is in charge of diagnosis, and the user is used to supply the system with observations. This architecture limits the application of diagnostic systems in areas such as healthcare, where highly skilled professional users do not readily accept a subordinate role. In addition, not taking users into account is one of the known causes of failure in knowledge-based systems (Brézillon, 1999). Finally, the cost of building and maintaining a diagnostic system based on AI techniques may very well be so large as to be impractical.

Alternatively, the diagnostic problem can be recast in terms of information access, rather than in terms of artificial intelligence. Complex AI architectures can be replaced by a collection of electronic texts that describe pathologies, which is searched by traditional techniques. Although this approach greatly simplifies knowledge base creation and maintenance, the limitations of information retrieval have been known for some time (Blair & Maron, 1985), and they indicate that locating the right information may be quite hard, and the result is usually not exhaustive.

More recently, systems based on hypertext technology have been proposed. An example is OncoDoc, a system for the assisted selection of clinical practice guidelines for cancer treatment (Bouaud et al., 1998; Séroussi et al., 2001). OncoDoc is an interactive system that encodes domain knowledge in the form of a decision tree implemented through a number of pages linked by hypertext links: the physician is presented with a sequence of choices that lead her to the guideline to be applied. Systems based on decision trees are less system-centric than conventional AI diagnostic systems, but user interaction is still quite rigid and follows predefined paths. Creation and maintenance are expensive, and the addition of a single new pathology may well disrupt the entire structure, and consequently user familiarity with the system.

DYNAMIC TAXONOMIES

In order to overcome the problems in current diagnostic systems, we propose a user-centric architecture in which the diagnostic system guides the user in exploring and systematically reduce the number of candidate pathologies until their number is sufficiently small for manual inspection. This architecture is based on a new information access paradigm represented by dynamic taxonomies (Sacco, 1987, 2000; also called *faceted search systems*), that focuses on exploratory, guided searches, rather than on traditional retrieval-based on precise specifications. Emphasis is placed on the user, on interactivity and on transparent operations, so that the main causes of failure in practical applications of diagnostic systems are removed. In addition, dynamic taxonomies easily support dynamic knowledge "encyclopedias," which include not only pathologies, but also emergency alerts, guidelines, and, in general, all relevant information. Since information items need not be textual, photos, x-rays, and so on can be easily accommodated.

Dynamic taxonomies are a general knowledge management model for complex, heterogeneous information bases. The intension is a taxonomy that does not require any other relationships in addition to subsumptions (e.g., IS-A and PART-OF relationships). Dynamic taxonomies require a multidimensional classification (i.e., items are classified under several topics at any level of abstraction, as appropriate). A concept C is just a label that identifies a set of items (items(C)), the set of the items classified directly under C, or under any of C's descendants (i.e., the *deep extension* of C).

This set-oriented approach has two important consequences. First, logical operations (and, or, not) on concepts can be performed by the corresponding set operations on their extension. Second, dynamic taxonomies can infer all the concepts related to a given concept C, which represent the conceptual summary of C. Concept relationships other than subsumptions are inferred by empirical evidence, through the extension only, according to the following *extensional inference rule*: two concepts A and B are related iff there is at least one item D in the infobase which is classified at the same time under A (or under one of A's descendants) and under B (or under one of B's descendants). For example, we can infer a (unnamed) relationship between Raphael and Rome, if an item that is classified under Michelangelo and Rome exists in the infobase. At the same time, since Rome is a descendant of Italy, also a relationship between Raphael and Italy can be inferred.

Dynamic taxonomies can be used to browse and explore the infobase in the following way. The user is initially presented with a tree representation of the initial taxonomy for the entire infobase. Each concept label has also a count of all the items classified under it (i.e., the cardinality of items(C) for all Cs). The initial user focus F is the universe (i.e., all the items in the infobase). In the simplest case, the user can then select a concept C in the taxonomy and zoom over it. This operation changes the current state as follows. First, concept C is used to refine the current focus F, which becomes $F \cap items(C)$; items not in the focus are discarded. Second, the tree representation of the taxonomy is modified in order to summarize the new focus. This means that all and only the concepts related to F are retained, and the count for each retained concept C' is updated to reflect the number of items in the focus F that are classified under C'. The reduced taxonomy is a conceptual summary of the set of documents identified by F, exactly as the original taxonomy is a conceptual summary of the universe.

The retrieval process can be seen as an iterative thinning of the information base: the user selects a focus, which restricts (thins out) the information base by discarding all the items not in the current focus. Only the concepts used to classify the items in the focus (and, because of subsumptions, their ancestors) are retained. These concepts, which summarize the current focus, are those and only those concepts that can be used for further refinements. From the human computer interaction point of view, the user is effectively guided to reach his goal, by a clear and consistent listing of all possible alternatives. Though guided, exploration is unconstrained because the user can freely focus on any concept, with the exception of those concepts not in the reduced taxonomy. These concepts are not related to the current focus, and focusing on any of them would produce a (useless) empty result.

The advantages of dynamic taxonomies over traditional access methods are dramatic in terms of convergence of exploratory patterns (Sacco, 2006), and in terms of human factors (Yee, 2003). Dynamic taxonomies have been successfully applied to a number of very diverse information access problems (Sacco, 2005a) and efficient implementations exist (Sacco, 1998).

DIAGNOSTIC ASSISTANCE THROUGH DYNAMIC TAXONOMIES

A dynamic taxonomy is used to recast the diagnostic problem in terms of guided information access to items that represent pathological situations, rather than in terms of logic manipulation. Each item is classified under one or more concepts that represent features that characterize it, including symptoms or causes of the pathology, but also locations in which a pathological situation is common, or age groups that exhibit it, and so on. In short, all the available information that can be used to characterize an item can be used as features. Features can be taxonomically organized in order to support generalization/specialization in accessing the knowledge base, and to simplify the selection of a specific feature when many features are present.

The diagnostic process is an interactive exploration of the knowledge base. Initially, the user is presented with the original taxonomy of features. From this taxonomy, he will zoom on the first symptom, say *fever*. The system will show a reduced taxonomy in which all and only those features (e.g., *chest pain*) related to high body temperature are retained. Related features are automatically computed on the basis of the extensional inference rule defined in the previous section: thus, *fever* and *chest pains* are related iff there is an item, such as pneumonia, in the knowledge base which is classified under both concepts. Related features can be used to further restrict the focus, and hence, the number of candidate items to be manually inspected. Focus refinement is obtained by AND'ing the previous focus with the selected new symptom/feature C.

Differently by traditional diagnosis, the goal is not to find the exact pathology, but rather a sufficiently small set of candidates that the user will finally and independently evaluate. Consequently, the diagnostic process can terminate at any time, and usually when the number of candidate items in the current focus is sufficiently small for manual inspection.

The reduced taxonomy indicates related features and symptoms, and clearly shows unexpected features that can be used to discriminate among items. This is especially important, since the practical importance of diagnostic systems is in dealing with the unknown and unexpected: a physician will hardly use a computer system to diagnose a flu; diagnosing a rare tropical disease that he has never seen nor read about is a completely different story (Sacco, 2005b).

DYNAMIC TAXONOMIES VS. DECISION TREES

There are superficial similarities between dynamic taxonomies and decision trees: both models use a decision structure to "drill down" and find appropriate items. In addition, after examining and selecting a value for a certain feature, only a subset of the remaining features needs to be tested. However, dynamic taxonomies are quite different from decision trees, both in information access (i.e., how the decision structure is used to identify an item) and in information classification (i.e., how the information structure is actually constructed).

The main difference in information access is that the user of a decision tree is presented with a fixed, predefined decision structure that he has to explore in the order defined by the designer (Moret, 1982). In current implementations such as OncoDoc, the decision tree must be followed in its entirety down to a leaf, with no indication of candidate items after each test.

Dynamic taxonomies, on the contrary, enforce no order among features, and the user can freely focus on any feature he finds relevant, in order to restrict the portion of the information base to be considered and to determine all the other features that can be tested, in a totally dynamic way. By selecting an arbitrary sequence of zoom operations, a specific decision tree out of all the possible ones is synthesized.

Such a freedom in selecting the sequence of tests has two important consequences. First, the sequence of tests can be tailored by the user, as to account for the actual cost of tests and to perform low-cost tests first, according to his specific environment. The cost of a test can depend on the specific capabilities of the medical unit: thus, a test that can be performed immediately in a central hospital might require the transportation of a patient from a decentralized unit. Therefore, any predefined sequence will be suboptimal in certain cases.

Second, selection by decision trees is the analog of interacting through a predefined modal dialog in which the system asks the questions: again, a system-centric master-slave architecture, although much more user-friendly than traditional AI systems. Access through dynamic taxonomies guides the user to reach his goal by simply hiding selections and features that do not apply in the current state produced by user selections. As such, interaction is modeless, and the system is actually the slave in the master-slave relationship, and is consequently perceived as an unobtrusive assistant.

At any stage, the user of a dynamic taxonomy system has a complete list of all the items that satisfied the tests performed so far. Consequently, the user can often rely on his knowledge/memory to select the right item in the candidates list, instead of applying additional tests. This can provide a significant speedup in locating appropriate items, especially when the corresponding decision tree is skewed. Finally, dynamic taxonomies support the notion of a taxonomic organization that guides users in easily locating available tests, and is especially important for all support material (guidelines, alerts, and so on).

As regards information classification, the notion of order required in decision trees is again critical. First, the total cost of tests must be minimized by anticipating low-cost tests, and deferring high-cost ones. As we argued above, some tests may have different costs in different environments, so that no fixed order is appropriate in all situations.

Second, the average or the maximum number of tests required should be minimized. Decision trees are usually incomplete binary trees, so that very skewed trees can arise in practice. Since the probability of selecting a specific information item is usually not uniform, it would make sense to select a tree that minimizes the number of tests for the most frequently occurring pathologies. But the frequency of pathologies depends on specific locations so that, in practice, the optimization of decision trees for diagnostic purposes is impossible.

The construction of a decision tree requires that n features that can be used to *discriminate* among items are identified, and then used to construct the decision structure. But, first, the addition of a new item requires, in general, that the existing decision tree be modified. In practical situations, modifications may well disrupt the entire structure, so that users have to relearn how to move in the structure.

Second, and more importantly, each item has to be compared against any other item in order to define appropriate discriminators. This results in a quadratic complexity in the number of items, and indicates that simple applications of decision trees do not easily scale up. By contrast, in dynamic taxonomies, features are used to *describe* each item in isolation. Consequently, the classification of a single item i depends on i only, and is therefore independent of the size of the information base.

CONCLUSION

Dynamic taxonomies are a promising approach for user-centered assisted diagnosis and decision support in healthcare. This model provides an interaction model that is guided and, at the same time, allows the user to freely focus on appropriate symptoms and features as desired. This allows to model the sequence of test required for diagnosis, as to account for local test costs and local frequency of pathologies. Current research focuses on building the taxonomy, possibly by automatic derivation from existing ontologies/thesauri such as UMLS (Lindberg et al., 1993) and SNOMED (2001). Other research directions include using electronic patient records for recommendation, and accounting for the probability of pathologies in guiding the search.

REFERENCES

Blair, D. C., Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Comm of the ACM, 28*(3), 289–299.

Bouaud, J., et al. (1998). Hypertextual navigation operationalizing generic clinical practice guidelines for patient-specific therapeutic decisions. In *Proceedings of AMIA Symposium* (pp. 488–492).

Brézillon, P. (1999). Context in problem solving: A survey. *The Knowledge Engineering Review, 14*(1), 1–34.

Buchanan, B. G., & Shortliffe, E.H. (1984). *Rule-based expert systems: The MYCIN experiments of the Stanford heuristic programming project*. Addison-Wesley, Reading, MA.

Console, L., Theseider Dupré, D., & Torasso, P. (1989). A theory of diagnosis for incomplete causal models. In *Proceedings of 11th International Joint Conference on Artificial* Intelligence (pp. 1311–1317). Detroit, MI.

de Kleer, J., Mackworth, A. K., & Reiter, R. (1992). Characterizing diagnoses and systems. *Artificial Intelligence, 52*, 197–222.

Lindberg, D. A. B., et al. (1993). *The unified medical language system*. *Med. Inform. Med., 32*, 281–291.

Moret, B. M. E. (1982). Decision trees and diagrams. *ACM Computing Surveys*, 14, 4.

Reggia, J. A., Nau, D. S., & Wang, Y. (1983). Diagnostic expert systems based on a set-covering model. *Internat. J. Man-Machine Studies, 19*, 437–460.

Sacco, G. M. (1987). Navigating the CD-ROM. In *Proceedings of International Conference Business of CD-ROM.*

Sacco, G.M. (1998). *Dynamic taxonomy process for browsing and retrieving information in large heterogeneous data bases*. U.S. Patent 6,763,349.

Sacco, G. M. (2000). Dynamic taxonomies: A model for large information bases. *IEEE Transactions on Knowledge and Data Engineering 12*(2), 468–479.

Sacco, G. M. (2005a). Dynamic taxonomies. In D. Schwartz (Ed.), *Encyclopedia of Knowledge Management*. Idea Group

Sacco, G. M. (2005b). Guided interactive diagnostic systems. *IEEE Symposium on Computer-Based Medical Systems.* Dublin.

Sacco, G. M. (2006, June). Analysis and validation of information access through mono, multidimensional and dynamic taxonomies. *FQAS 2006, 7th International Conference on Flexible Query Answering Systems.* Milano. Springer LNAI 4027.

Séroussi, B., et al. (2001). OncoDoc, a successful experiment of computer-supported guideline development and implementation in the treatment of breast cancer. *Artif Intell Med, 22*(1), 43–64.

SNOMED. (2001). SNOMED International: Clinical terms. Retrieved February 6, 2008, from http://www.snomed.org

Wollersheim, D., & Rahayu, W. (2002). Methodology for creating a sample subset of dynamic taxonomy to use in navigating medical text databases. In *Proceedings of IDEAS 2002 Conference*. Edmonton, Canada.

Yee, K. -P., et al. (2003). Faceted metadata for image search and browsing. In *Proceedings of ACM CHI 2003*.

KEY TERMS

Decision Tree: A rooted, ordered, and labeled tree. Given a boolean function on n-variables (tests), a decision tree is a deterministic algorithm for deciding which variable to test next, based on the previously tested variables and the results of their evaluation, until the function's value can be determined.

Deep Extension: Of a concept C, denotes the shallow extension of C union the deep extension of C's sons.

Extensional Inference Rule: Two concepts A and B are related iff there is at least one item *d* in the knowledge base which is classified at the same time under A (or under one of A's descendants) and under B (or under one of B's descendants).

Facet: One of several top-level (most general) concepts in a multidimensional taxonomy. In general, facets are independent, and define a set of "orthogonal" conceptual coordinates.

Monodimensional Taxonomy: Taxonomy where an item can be classified under a single concept only.

Multidimensional Taxonomy: Taxonomy where an item can be classified under several concepts.

Reduced Taxonomy: In a dynamic taxonomy, a taxonomy, describing the current user focus set F, which is derived from the original taxonomy by pruning from it all the concepts not related to F.

Shallow Extension: Of a concept C, denotes the set of documents classified directly under C.

Subsumption: A subsumes B if the set denoted by B is a subset of the set denoted by A ($B \subseteq A$).

Taxonomy: A hierarchical organization of concepts going from the most general (topmost), to the most specific concepts. A taxonomy supports abstraction and models subsumption (e.g., IS-A and/or PART-OF) relations between a concept and its father.

User Focus: The set of documents corresponding to a user-defined composition of concepts; initially, the entire knowledge base.

Zoom: A user interface operation that defines a new user focus by OR'ing user-selected concepts and AND'ing them with the previous focus; a reduced taxonomy is then computed and shown to the user.

Haptic Rendering for Laparoscopic Surgery Simulation and Related Studies

Ryan McColl
Monash University, Australia

Ian Brown
Monash University, Australia

Cory Seligman
Monash University, Australia

Fabian Lim
Monash University, Australia

Amer Alsaraira
Monash University, Australia

INTRODUCTION

This project concerns the application of haptic feedback to a virtual reality laparoscopic surgery simulator. It investigates the hardware required to display haptic forces, and the software required to generate realistic and stable haptic properties. A number of surgery-based studies are undertaken using the developed haptic device.

The human sense of touch, or haptic sensory system, is investigated in the context of laparoscopic surgery, where the long laparoscopic instruments reduce haptic sensation. Nonetheless, the sense of touch plays a vital role in navigation, palpation, cutting, tissue manipulation, and pathology detection in surgery. The overall haptic effect has been decomposed into a finite number of haptic attributes. The haptic attributes of mass, friction, stiction, elasticity, and viscosity are individually modeled, validated, and applied to virtual anatomical objects in visual simulations.

There are times in surgery when the view from the camera cannot be depended upon. When visual feedback is impeded, haptic feedback must be relied upon more by the surgeon. A realistic simulator should include some sort of visual impedance. Results from a simple tissue holding task suggested the inclusion of haptic feedback in a simulator aids the user when visual feedback is impeded.

BACKGROUND

Introduction to Haptic Feedback

Haptic perception is equivalent to humans' sense of touch, and incorporates both tactile and kinesthetic perception. Haptic feedback is present whenever the haptic receptors are engaged. All physical objects intrinsically display haptic feedback to humans when they interact with them, and include texture, weight, and heat. In relation to this document and simulators in general, haptic feedback is the term given to the forces displayed by a mechanical haptic device able to simulate an objects haptic information.

Surgery simulation and robotic surgery is an area where haptic feedback use is increasing. The sense of touch provides useful information to surgeons when orientating themselves, diagnosing pathologies, and manipulating tissue. Visual-only simulators can improve surgical training and education, but hands-on experience can only be achieved by including haptic feedback (Nudehi, Mukherjee, & Ghodoussi, 2003, p. 4563). To ensure a completely immersive and realistic VR simulator, haptic feedback must be included. Due to computing power issues in the past, the implementation and development of a haptic feedback system has come secondary to the highly computational visual graphic components. Improvements in technology and further research will lead to haptic feedback becoming more reliable, stable, and commonplace.

Haptic Perception

The sense of touch in human is extremely important and often underrated (Robles-De-La-Torre, 2006). A realistic surgery simulator must provide sufficient visual and haptic information to the user. Simulated haptics and graphics is not real life, but by exploiting the limitations of humans' senses, realism is created. To synthesise haptic feedback to such a level, an understanding of the biomechanical, sensorimotor, and cognitive abilities of the human is required (Tan, Srinivasan, Eberman, & Cheng, 1994, p. 353).

One way of testing and comparing humans' abilities is to use the Just-Noticeable-Difference (JND) measure. The Just-Noticeable-Difference (JND) describes the smallest amount of change of an arbitrary stimulus which can be detected, just as often as it cannot be detected. Therefore, the JND, in the context of force perception, is defined as being the minimum amount of force change to an initial force that a user can detect in 50% of trials (Wikipedia, 2006). It can be used to test many haptic attributes, and provides a means to validate hardware and software capabilities in a haptic system. This value lies mostly between 5% and 15% for humans, depending on the stimulus type and limb or joint where it is applied (Hinterseer & Steinbach, 2006).

Sense Perception During Laparoscopic Surgery

There is significant degradation and distortion of forces encountered by surgeons during laparoscopic procedures. The force feedback from tissue interaction, including tasks such as tissue palpation, is effected by friction at the trocar, weight of the instrument, and length of the instrument (Patkin & Isabel, 1995; Tavakoli, Patel, & Moallem, 2006). As a reference, a metal trocar can introduce about 0.4N of friction (Sallé, Gosseli, Bidaud, & Gravez, 2001), and the weight and length of a typical laparoscopic instrument are about 100grams (1N) and 400mm respectively.

Depth perception cues are lost, due to viewing the scene on a 2D monitor. Patkin and Isabel (1995) found that five out of 16 depth perception cues are lost including stereopsis, convergence, and texture fade. Also, the field of view reduces from 150° to 60°, and viewing distance reduces by 90%, compared to open surgery. Hand-eye coordination is reduced by the indirect method of observing and manipulating tissue. Breedveld and Wentink (2001, p. 155) suggest this is due to a mislocation of the camera picture, and a misorientation of the instrument movements. The instruments movement on the monitor are different than expected, and the "effects are very confusing." They, however, make no mention on the role force feedback plays in compensating for the hand-eye coordination effect. Bingham et al. (2000) does consider haptic feedback in a hand-eye coordination experiment. It does not directly relate to laparoscopic surgery, but results suggest that distance and shape distortion created by visual illusions can be corrected somewhat by haptic feedback.

Training simulators should include haptic feedback to improve users' awareness of its role. Development of a surgeons' sense of touch can reduce damage to organs and tissues, and improve recognition of the slightest difference of tissue density, improving the ability to identify blood vessels and other structures merely by touch alone (Patkin, 2005; Patkin et al., 1995).

Haptic Hardware

An ideal mechanical interface acts as a transparent medium in which force is displayed to the user from the data calculated in the virtual anatomy model. It should be friction free, have zero mass and inertia, have zero backlash, and display a wide magnitude of forces without error. This ideal will never be realised in the real world, so we strive for a device as close to ideal as possible.

Haptic Modeling

Surgery simulation requires a force reflecting deformable model of the anatomy. The quality of visual and haptic information must be sufficient to simulate a high level of realism to the user. Accurate physical-based models are computationally demanding, whereas simple geometric models do not generate realistic information. A balance between realism and real-time operation is required.

Validation

In order to determine the effectiveness of various haptic models developed within this project, validation studies need to be undertaken. The hope is that haptics enable surgeons to improve skills faster than if haptics were not included. However, there are two other scenarios, being that haptics may achieve no improvements, or may actually hinder the surgeon's learning. Therefore, validation studies must be conducted to justify the inclusion and measure the quality of the haptics.

SYSTEM OVERVIEW

Figure 1 provides an overview of the mechanical interface setup for a two-instrument and one-camera task. The mechanical interface provides a two-way flow of haptic and positional information between the users hands and the virtual reality world. Positional information is fed via positional encoders to the PC, and haptic information is fed via the PC, through the motor amplifiers, motors, and gimbal, up to the instrument handles. Information flow occurs simultaneously in 3D for each instrument.

Haptic Hardware

The mechanical interface was designed and built by the authors (McColl, Brown, Seligman, Lim, & Alsaraira, 2005; Seligman, 2003). Gears have been used to multiply torque transfer for parallel transmission, ensuring the required display of continuous 6N forces to the instrument handles are met. A pulley and belt provides smooth torque transfer, but the belt tension to drive required forces without slippage caused mechanical distortion. Gears are nonslip, but have inherent backlash. The backlash is not noticeable in normal operation. Figure 2 shows gimbal movement and geared torque transmission setup.

A virtual hemisphere study was performed to compare subjects' ability to locate the apex of a virtual hemisphere with the haptic device under maximum force capabilities of 2N, and then 6N. The 6N mean error was 21.7% with standard deviation 13.5. This suggests a marked improvement from the 2N results (mean error 29.7%, standard deviation 17.3). The navigational results found by increasing available force output improves navigational ability, and corresponds with findings from O'Malley and Goldfarb (2002).

The Faulhaber (Faulhaber Group, http://www.faulhaber-group.com) ironless core DC motor with graphite commutation (Model Number: 3863-048C) was chosen for the larger parallel motors. The serial motors used are Faulhaber (Faulhaber Group) ironless core DC motors with precious metal commutation (Model Number: 2224-006SR). They are much smaller than the parallel motors, and only generate about 1.5% of the torque in comparison. The light weight is essential as the motor moves along the pitch and heading axes with the instrument. It directly contributes to the device inertia. A capstan with a small moment arm was designed to maximise the force output from the small torque capability.

The brush DC motors were chosen for their low inertia, low stall torque, ironless core, no magnetic cogging, and low axial play. Other options for motors include brushless and stepper motors. Brushless motors require extra control circuitry and complexity, and suffer from reluctance cogging and torque ripple phenomena (Tavakoli et al., 2006, p. 164). Stepper motors are noisier, have greater torque ripple, and also require

Figure 1. Overview of haptic interface

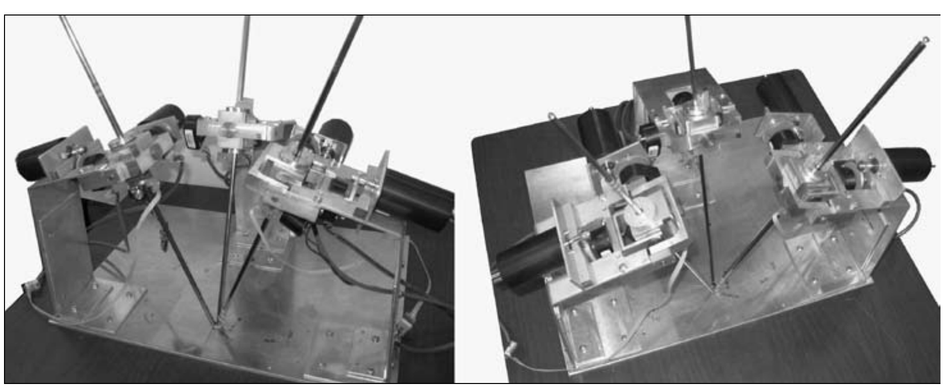
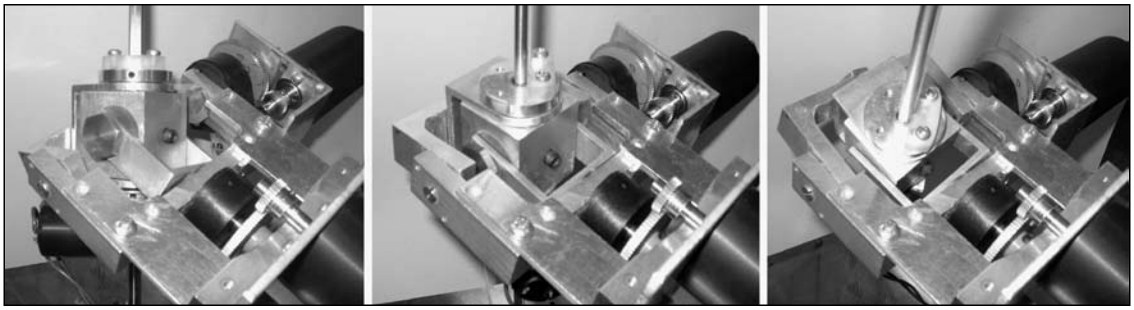


Figure 2. Gimbal movement and geared torque transmission



more complex circuitry, in comparison with brush type DC motors (Chen & Marcus, 1998, p. 526).

Haptic Software

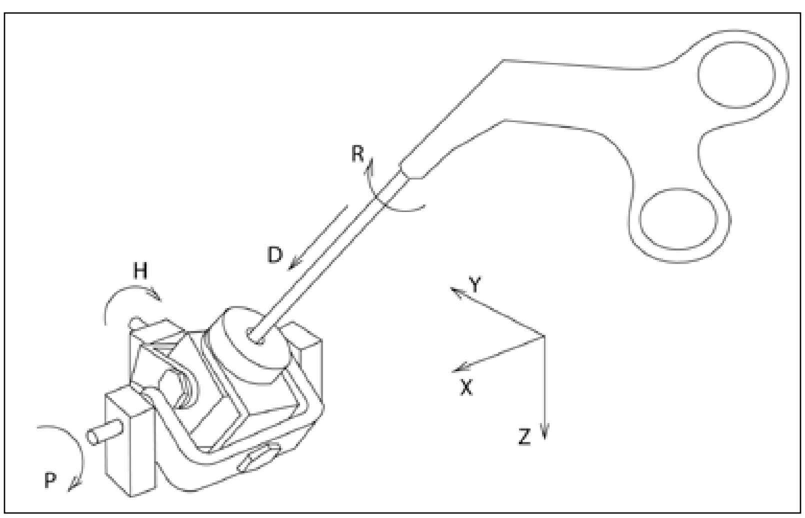
The gimbal allows instrument movement in four degrees of freedom (H,P,R,D). In the current hardware implementation, the H axis (heading) reflects rotation of the instrument left and right, the P axis (pitch) represents rotation forward and back, the R axis (roll) represents rotation of the handle about its own axis, and the D axis (depth) represents axial movement of the instrument through the centre of the gimbal. Cartesian coordinates (X,Y,Z) are used by the visual loop, and relationships must be made in order to pass information between visual and haptic loops. The relationship between polar and Cartesian coordinates for the gimbal is shown below in Figure 3.

A Cartesian normal surface vector is passed from the visual loop when contact is made between an instrument and an object. This is converted to a force vector based on attributes of the organ, such as mass or deformation. In order to display this force to the user through the mechanical interface, the Cartesian vector must be separated into polar coordinates, which allows the signals for the force feedback motors to be generated.

Haptic attributes are decomposed properties of the overall haptic effect. It is part of the haptic modeling process to determine how to decompose the overall effect into attributes that can be modelled, and then rendered onto 3D deformable objects. This project provides a unique decomposition which is suitable for laparoscopic surgery simulation.

The overall haptic effect can be effectively created by the summation of the individual attributes. Each virtual anatomical object is defined a set of attributes

Figure 3. Coordinate system



which sum together, creating an overall force. Attributes, which in programming terms can be thought of as objects, can be added, subtracted, and reprogrammed without affecting other attributes. This nature of this object-oriented approach model lends itself neatly to the addition of new objects and features into the VR model.

The process of decomposing the overall complex haptic effect into simple attributes began with discussion with a medical consultation group (Healy & Rombauts, 2005), and observation of real laparoscopic surgical procedures in theatre and on video. The technique involved generation of a potential attribute list, programming potential attributes, parameter adjusting, algorithm modification, subjective evaluation, and objective validation.

The overall haptic effect has been decomposed into five attributes: mass, elasticity, friction, stiction, and viscosity. This is not a definitive decomposition of the overall haptic effect.

The haptic effect is too complex to expect that the breakdown here can satisfy an entirely realistic response. This list is one solution to a haptic modeling problem, and is believed to be a good method for this application.

Forces relating to mass stem from gravitational effects and acceleration. Elasticity is a combination of a stiffness and a damping component. Friction, which is used for simulating texture and cutting, is modeled simply using modified random noise functions. Parameters for the attributes have been subjectively tuned using real tissue properties as baselines.

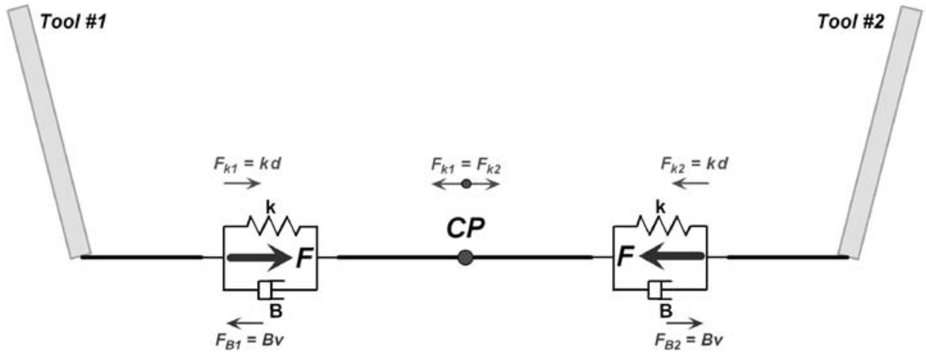
A representation of both viscosity and stiction has been achieved using the following algorithm. Stiction can be thought of as any force which opposes motion along the surface of an object. An example is a diathermy tool sticking to tissue or resistance, due to viscous surface fluid.

- When contact is made between the Tool Tip (*TT*) and an object, store the Point of Contact (*POC*). Note that inside viscous liquid, the *TT* is always touching an object.
- As the *TT* moves away from the *POC*, apply a force to the *TT* in the direction of the *POC,* based on the distance between the two points.
- Whilst the *TT* remains in contact with the object, the *POC* moves towards the *TT* at a speed based on the distance between the two points.
- Once the *TT* is no longer geometrically touching the object, a force is maintained for a length of time based on distance between the *POC*. This ensures there remains some physical contact, or stiction.
- Properties can be varied greatly by adjusting parameters related to distance, speed, and time.

The haptic loop obtains information from the virtual world via the visual loop. The geometric visual models are rendered with one or more of the haptic attributes. This, along with some simple passivity algorithms, adds a physical element to the geometric models.

Multiple instruments acting upon the same object has historically been a challenge and a cause of instability. Presented here is a simple approach which combines the modular concept of Yoshikawa and Ueda (1996), and the simple spring-damper model used by Basdogan, Ho, Srinivasan, and Slater (2000). The simplest case is an elastic ligament held by two instruments. The ligament has a stiffness and damping component, but displays no mass. Forces are generated only when the ligament is stretched beyond its nominal (or resting) length. The stiffness component generates forces for both instruments, equal in magnitude but in opposite directions. The direction of the force vector points directly towards the centre point (CP), which in this case is the same direction as the opposing tool tip. The damping component acts along the same vector direction as the force component, but the magnitudes of each instrument are not necessarily the same. The damping magnitude is based on the velocity each individual instrument is travelling. An ideal damper, as used for this ligament model, absorbs energy, and does not transmit force through itself. Each instrument acts independently, not requiring knowledge about the other instrument. The only shared data is the CP location and the ligament properties. A visual representation is shown in Figure 4. This concept can be used for more complex objects by adding ligaments attached to fixed locations or part of an object. The simulation displays accurate and stable forces for one-person (one-handed and two-handed) operation, and two-person operation.

Figure 4. Multiple instrument ligament interaction



EXPERIMENTAL STUDIES

Threshold Perception Study

- **Aim:** To validate the haptic attributes by measuring the JND for each attribute. This ensures both that forces that cannot be perceived are not unnecessarily displayed, and that our mechanical interface accurately reproduces forces to users, and is suitable for further studies investigating surgery-based haptic tasks.
- **Overview:** Studies have shown humans can detect change in forces of approximately 10% through their fingers (Allin, Matsuoka, & Klatzky, 2002; Doerrer & Werthschuetzky, 2002; Pang, Tan, & Durlach, 1991). The validation of subjects' ability to detect changes in haptic attributes is more difficult to quantitatively measure, as equations used to model attributes vary from the physical world. A simple way to achieve validation is to determine what the JND for force is through the changing of parameters in the attribute equations. For example, what were the forces involved in a subject detecting a change in viscosity of 10%?
- **Method:** The following perception studies were undertaken with each subject: mass, tissue elasticity, surface stiction, liquid viscosity, and detection of an arterial pulse (displacement). Following familiarisation and a trial run, subjects were tested 30 times for each attribute, 10 times for each of three default levels. The increment and direction of the changes were randomly generated. A computer-generated tone sounded each time the attribute changed. Subjects were then given a chance to determine if a change had occurred. Five male subjects aged 20–35 in total were tested.
- **Results:** Average JND for mass was 12.5%. The smallest change detectable from a zero base was 0.13N. The instrument weight of 1N makes JND 13% about a zero base. For base forces of 1.25N and 2N, JND was 12%.

The JND for an elastic membrane with stiffness of 100N/m and 500N/m was 10%, and for 1500N/m, was 16.7%.

The average JND for stiction was 13%. A change in stiction parameters is proportional to a change in force for a given velocity. Each subject has control over their velocity during this experiment, so the JND is based on both force and velocity.

The average JND for viscosity was 8%. As with the stiction calculations, force change is proportional to viscosity change, but forces generated depend on the velocity of the instrument.

Subjects were able to detect, on average, 5% changes in the height of a simulated arterial pulse. The arterial pulse is modeled as a height changing membrane with stiffness of 1000N/m. Perception is dependent on users' ability to detect both a change in force and change in displacement.

The overall JND for purely force related tests is approximately 12%. This suggests our interface is suitable for use with surgery-based tasks. The higher JND for the high stiffness membrane suggests the mechanical limitations of the interface may be being approached. The remaining attributes tested depend on force perception, and either velocity or displacement perception. Further studies are required to comprehend further, but low percentage JND's suggest our modeling techniques provide measurable information to users, which can be applied to deformable visual graphic models.

Object Localisation Study

- **Aim:** To test the value of various modes of haptic/visual feedback in a simulated laparoscopic surgery task. It investigates localisation of an anatomical object to a spatially known location. Several studies, for example Jones, Bokinsky, Tretter, and Negishi (2005) and Moody, Baber, and Arvanitis (2001) investigate performance tasks with constant and continual visual feedback. This study investigates how performance of a task is affected as visual feedback is impeded slightly or removed.
- **Overview:** Subjects are given an opportunity to initially localise the instrument to the desired position using an active visual graphic position monitor. The active monitor is removed, and subjects must then rely on the modes listed below to relocate the instrument to the desired position. 10 male subjects aged 20-35 in total were tested.
- **Modes of visual/haptic feedback:**
  - None (limb localisation feedback only) (NF)
  - Fixed Viewpoint Visual Feedback (FVF)
  - Haptic Feedback (HF)
  - Fixed Viewpoint Visual & Haptic Feedback (FV&HF)
  - Varying Viewpoint Visual Feedback (VVF)
  - Varying Viewpoint Visual & Haptic Feedback (VV&HF)

  Visual feedback displays a virtual view of the instrument inside the virtual abdomen. The camera viewpoint is located to approximate the position of a real camera during surgery.

  The FVF displays the scene from a fixed viewpoint. Subjects can take advantage of this by visually lining up the instrument with tokens on the monitor.

  The VVF displays the scene from a camera randomly drifting in three axes. Observation of surgery suggests this is more realistic than a FV model. Position and angle drift is limited to approximately ±10% from the initial position, and can move between the limits in approximately 10 seconds.

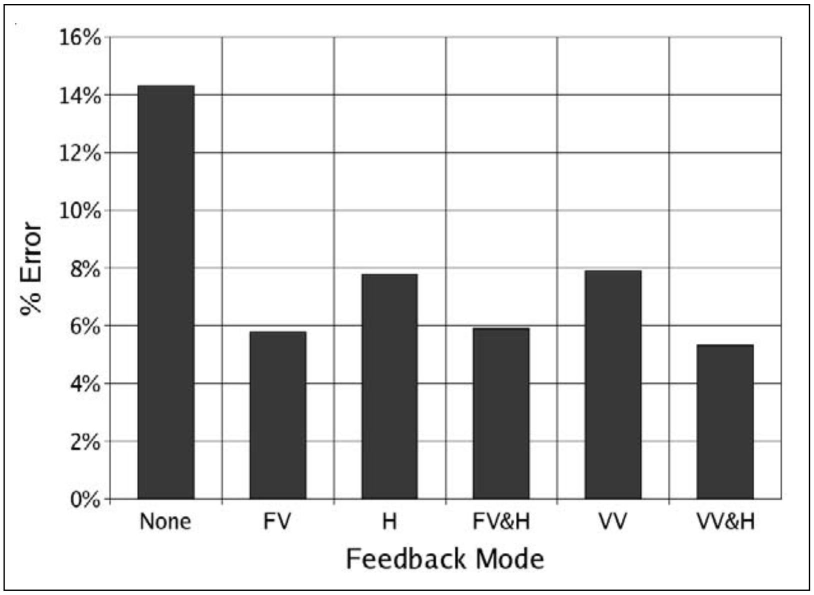
  HF displays a force to the user without any visual feedback. A simple elastic model is used for the object being stretched. The desired position represents a movement of 60mm from the initial position, generating a force of 2.5N.

  NF indicates no visual or haptic feedback. Subjects' eyes are covered, and haptic feedback is switched off. The ability to locate the instrument is based purely on the ability of the subject to localise themselves using the sensed angle of their shoulder, elbow, and wrist joints. This provides a benchmark for other modes.
- **Method:** Following familiarisation and a trial run, subjects perform the localisation task five times with each feedback mode. After the five attempts, subjects stand, walk around, then localise themselves again using the active position monitor under the new feedback mode conditions. The whole experiment is performed twice per subject.
- **Results:** The percentage error in locating to the desired position without visual or haptic feedback is approximately 15%. With FVF only the percentage error in locating is approximately 6%. With VVF, only the percentage error in locating is approximately 8%. The difference makes sense, as it is more difficult to locate something using vision if your frame of reference is changing. Haptics alone measured a percentage error in locating of approximately 8%.

The most interesting results combine visual and haptic feedback. With FV&HF the percentage error in locating is still approximately 6%. This suggests haptics neither detracts nor adds to the ability to locate under these circumstances when the visual feedback is fixed. With VV&HF the percentage error is approximately 6%, a 2% improvement on both VVF & HF alone. This suggests haptics aids in the ability to locate an object when the visual viewpoint is changing.

A plot of results can be found in Figure 5. To analyse the data, an ANOVA with three factors, Feedback(FB) Mode, Subject, and Trial, was performed. Subject was considered a random factor. Differences between FB-Modes and Trial were found to be highly significant ($p<0.001$ and $p=0.002$). Tukey's post-hoc test split the

Figure 5. Object localisation results



FBModes into three groups: NF; FV, FV&H, VV&H; and H, VV. (The assumptions of ANOVA were met.) Interaction plots suggested that the effect of Trial was chiefly due to the difference between Trials 1 and 2 for NF. When NF was removed, the analysis no longer found Trial significant, but Subject was now significant (p=0.012), because the MS error for Trials had diminished by a factor of 5. This suggests that even though each Subject had a different haptic ability, their comparative ability across FBModes showed significant trends. The groupings determined from Tukey's post-hoc test suggest that the addition of haptic feedback, when visual feedback is impeded, improves the performance of this task.

FUTURE TRENDS

The haptic community will continue to grow, providing more people with access to haptic devices. Surgery simulation will continue to be an area heavily involved in haptics. It appears that hospitals and staff are ready to take on new products, but difficulty of use, unreliability, nonrealistic haptic sensations, and pretraining requirements, will hinder their uptake.

More studies investigating how haptics effect performance will continue to generate knowledge on how the haptic sense works, and more importantly in this context, how to best develop training procedures or warm-up routines to best focus attention to the development of important skills required for surgery simulation.

Though several commercial devices exist, it is important that researchers continue work on all aspects of haptic systems. Commercialisation often limits common knowledge with black box components (Çavuşoğlu, Feygin, & Tendick, 2002, p. 556), closed source software and patents, hindering the ability for new users to learn how to best develop their own for a custom application.

CONCLUSION

Studies are required to validate new haptic hardware and software to ensure its suitability. Our mechanical interface has been validated using a virtual hemisphere navigation study, and a perception study measuring the JND of forces displayed to the user. Haptic attribute models operate at real-time, and provide measurable information to users. We believe our system is suitable for further studies in the context of surgery simulation.

A tissue localisation study was conducted, which compared performance using different modes of feedback. Results suggested that including haptic information to subjects improved their ability to locate tissue to a known position when visual feedback is impeded.

The development of haptic rendering hardware and software has identified a number of important engineering design challenges. Solutions to these challenges need to be both cost-effective and computationally efficient. Global physical modeling may produce a single virtual organ with realistic visual and haptic properties, but a multiorgan, multiinstrument, and multioperator simulation suggests a less global, more segmented, object-oriented approach. This approach is currently being investigated in relation to the haptic rendering of a virtual reality laparoscopic simulator.

REFERENCES

Allin, S., Matsuoka, Y., & Klatzky, R. (2002, April). Measuring just noticeable differences for haptic force feedback: Implications for rehabilitation. *IEEE Haptics Interfaces for Virtual Envir. & Teleoperator Systs.*

Basdogan, C., Ho, C. -H., Srinivasan, M., & Slater, M. (2000, December). An experimental study on the role of touch in shared virtual environments. In *ACM Trans. Comput.-Hum. Interact.* (Vol. 7, No. 4., pp. 443–460). ACM Press.

Bingham, G.P., Zaal, F., Robin, D., & Shull, J.A. (2000). Distortions in definite distance and shape perception as measured by reaching without and with haptic feedback. *Journal of Experimental Psychology: Human Perception and Performance*, 26(4), 1436-1460.

Breedveld, P., & Wentink, M. (2001, May). Eye-hand coordination in laparoscopy - an overview of experiments and supporting aids. *Minimally Invasive Therapy & Allied Technologies*, *10*(3), 155–162.

Çavuşoğlu, M., Feygin, D., & Tendick, F. (2002). Critical study of the mechanical and electrical properties of the PHANToM haptic interface and improvements for high performance control. *Presence: Teleoperators and Virtual Environments*, *11*(6), 555–568.

Chen, E., & Marcus, B. (1998, March). Force feedback for surgical simulation. In *Proceedings ofthe IEEE* (Vol. 86, No. 3, pp. 524–530).

Doerrer, C., & Werthschuetzky, R. (2002). Simulating push-buttons with a haptic display: Requirements of force resolution and force-displacement curve. In *Proceedings of EuroHaptics 2002*. Edinburgh, UK.

Haans, A., & Ijsselsteijn, W. (2006, March). Mediated social touch: A review of current research and future directions. *Virtual Reality*, *9*(2-3), 149–159.

Healy, D., & Rombauts, L. (2005-6). *Monash medical gynaecological surgeons*. Private Communication.

Hinterseer, P., & Steinbach, E. (2006, March). A psychophysically motivated compression approach for 3D haptic data. *Symposium on Haptic Interfaces for Virtual Environments and Teleoperator Systems*. Alexandria, Virginia, USA

Jones, M., Bokinsky, A., Tretter, T., & Negishi, A. (2005, May). A comparison of learning with haptic and visual modalities. *Haptics-e, The Electronic Journal of Haptics Research*, *3*(6). Retrieved February 6, 2008, from http://www.haptics-e.org

McColl, R., Brown, I., Seligman, C., Lim, F., & Alsaraira, A. (2005). Haptic rendering for VR laparoscopic surgery simulation. *Engineering and the Physical Sciences in Medicine.* Adelaide, Australia.

Moody, L., Baber, C., & Arvanitis, T. (2001). The role of haptic feedback in the training and assessment of surgeons using a virtual environment. In *Proceedings of Eurohaptics 2001.* Univeristy of Birmingham, UK.

Nudehi, S., Mukherjee, R., & Ghodoussi, M. (2003, December). A haptic interface design for minimally invasive telesurgical training and collaboration in the presence of time delay. In *Proceedings of 42nd IEEE Conference on Decision and Control* (pp. 4563–4568). Maui, HI.

O'Malley, M. K., & Goldfarb, M. (2002, September). The effect of force saturation on the haptic perception of detail. *IEEE/ASME Transactions on Mechatronics*, *7*(3), 280–288.

Pang, X. -D., Tan, H. Z., & Durlach, N. I. (1991). Manual discrimination of force using active finger motion. *Perception & Psychophysics, 49*(6), 531–540.

Patkin, M. (2005). *Helping at lap surgery*. Retrieved February 6, 2008, from http://www.mpatkin.org/surg lap/lap_helping1.htm

Patkin, M., & Isabel, L. (1995). Ergonomics, engineering, and surgery dissection. *J. R. Coll. Surg. Edin.*, 40, 120–132.

Robles-De-La-Torre, G. (2006, July/September). The importance of the sense of touch in virtual and real environments. *IEEE Trans. Multimedia*, *13*(3), 24–30.

Salisbury, K., Conti, F., & Barbagli, F. (2004, March/April). Haptic rendering: Introductory concepts. *IEEE Comput. Graph. Appl*., *24*(2), 24–32.

Sallé, D., Gosseli, F., Bidaud, P., & Gravez, P. (2001, September 18–21). Analysis of haptic feedback performaces in telesurgery robotic systems. In *10th IEEE International Workshop on Robot-Human Interactive Communication* (pp. 618–623). Bordeaux and Paris, France.

Seligman, C. (2003). *Force feedback haptics for virtual reality laparoscopic surgery*. Unpublished master's thesis, Monash University, Department of Electrical and Computer Systems Engineering, Clayton, Australia.

Tan, H., Srinivasan, M., Eberman, B., & Cheng, B. (1994, November). Human factors for the design of force-reflecting haptic interfaces. In *Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems* (pp. 353–359). Chicago, IL.

Tavakoli, M., Patel, R., & Moallem, M. (2006, March). A haptic interface for computer-integrated endoscopic surgery and training. *Virtual Reality, 9*(2–3), 160–176.

Wikipedia. (2006). *Just noticeable difference.* Retrieved February 6, 2008, from http://en.wikipedia.org/w/index.php?title=Just_noticeable_ difference&oldid=85626600

Yoshikawa, T., & Ueda, H. (1996, April 22–28). Construction of virtual world using dynamics modules and interaction modules. In *Proceedings of IEEE International Conference on Robotics and Automation* (Vol. 3, pp. 2358–2364).

KEY TERMS

Force Feedback or Kinesthetic Feedback: Refers to the information received via receptors in the limbs themselves. It is responsible for awareness of movement and position of one's limbs in space and time. Kinesthetic receptors are also responsible for the detection of size and weight of objects (Haans & Ijsselsteijn, 2006).

Haptic Attributes: All physical objects intrinsically display haptic attributes which feedback haptic information to humans interacting with them. The haptic attributes include texture, heat, mass, friction, stiction, elasticity, viscosity, temperature, vibrations, and more.

Haptic Attribute Decomposition: The overall haptic effect can be decomposed into a number of haptic attributes. The decomposition attaints to determine the method in which best to decompose the overall haptic effect. Once the decomposition has occurred, the individual attributes can be measured, compared, and recreated in a simulation.

Haptic Display or Force Display: Refers to the transfer of force information via the mechanical interface to the user. The case of a force display is analogous to that of a visual display, where visual information is passed to a user via a medium such as a computer monitor.

Haptic Rendering: Haptic rendering is the process of converting computer algorithms containing force information to a mechanical interface capable of displaying haptic information to a user (Salisbury et al., 2004). It is analogous to visual rendering, which converts graphics from a file into visual form, as on a video display. Haptic rendering is comprised of a hardware and a software component.

Haptics: From the Greek word *haptesthai*, meaning "to touch." Haptics broadly refers to touch interactions (physical contact) that occur for the purpose of perception or manipulation of objects (Salisbury, Conti, & Barbagli, 2004). This encapsulates both "Tactile" and "Force" information.

Tactile Feedback or Cutaneous Feedback: Encompasses all the information acquired through sensors in the skin, with particular reference to the spatial distribution of pressure, or more generally, tractions, across the contact area. Tactile Feedback is responsible for the detection of, but not limited to, roughness, temperature, and vibrations.

Visual Impedance: "Impedance" can be defined as something that impedes, such as an obstacle or hindrance. Therefore, a visual impedance is something that impedes vision.

Health Systems Simulation

David Lyell
University of New South Wales, Australia

Rosemarie Sadsad
University of New South Wales, Australia

Andrew Georgiou
The University of Sydney, Australia

INTRODUCTION

The health system is the culmination of organisations, institutions, and resources which work together for the purpose of providing health care and improvement in health (WHO, 2006). This system comprises a number of different levels (Proctor, Reid, Compton, Grossman, & Fanjiang, 2005) as shown in Figure 1. The highest level is the environment which provides the regulatory, policy, and funding frameworks in which the lower levels of the health system operate.

The health system is a *complex system* (Plsek & Greenhalgh, 2001) meaning that it is highly coupled and many outcomes are the result of the interactions that occur between many different parts (Plsek, 2003). Not only can interactions occur between different parts at the same level, but they can also interact across levels.

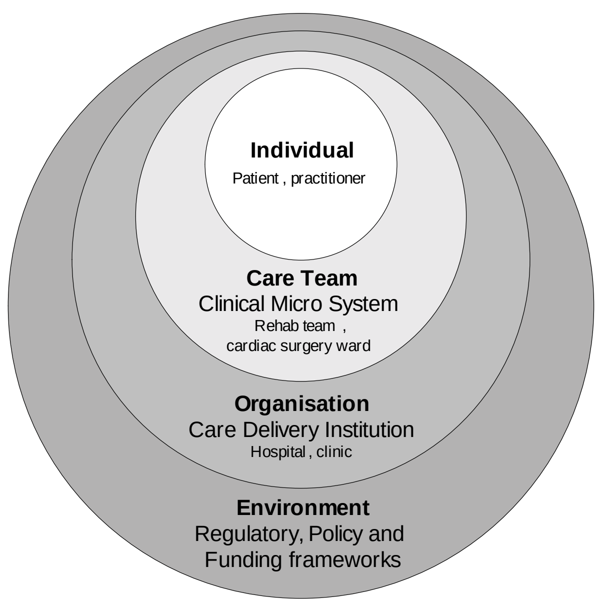
Adding to the complexity of the health system is that these interactions play out over time, often with effects being distant in both time and place to causes (Forrester, 1961). This *dynamic complexity* (Sterman, 2000) arises when systems change over time, are tightly coupled (characterised by strong interactions), are governed by feedback, and play out over time. Research has shown that people are very poor at understanding problems with dynamic complexity, with performance on tasks and outcomes declining as complexity or the strength of feedback increases (Paich & Sterman, 1993).

In this way, the health system is more than simply the sum of its parts because all aspects of the system need to function together in order to provide health care. One of the main difficulties in health care reform therefore is the systemic nature of health care. Reform to one area will undoubtedly affect other areas.

Most problems arising from the operation of the health system are studied and addressed using conventional reductionist methods, which reduce, isolate, and freeze aspects of the system at a given time. This fails to deal with the dynamic complexity inherent in the health system and which is often the source of the problem. The result is that all too often, well intentioned interventions make the original problem worse by failing to fully understand the complexities involved in the origin of a problem (Sterman, 2000).

In this article, we introduce *system simulation* as a means of exposing the underlying causes and systemic structures of problems within the health care system, as well as providing a tool for assessing the likely impact of new interventions. The following sections will examine the advantages of simulation, areas of application, how simulation experiments can overcome some of the limitations of randomised control trials (RCTs), and various simulation methodologies as well as the challenges of conducting simulation experiments.

Figure 1. Conceptual drawing of the 4 levels health care system (Adapted from Proctor et al., 2005)



HEALTH SYSTEM SIMULATION

Health system simulation is the application of modelling and computer simulation methods to study the interactions between individuals and/or components of a system and how these interactions over time produce the behaviour observed in the health care system.

There are both quantitative and qualitative modelling and simulation methods, however this article will focus exclusively on quantitative modelling and simulation. See Kuipers (1994) for a discussion on qualitative simulation methods.

Health system simulation helps develop an understanding of the dynamic relationships and systemic causes of problems. Interventions can be developed to address these problems and can be tested *in silico* to determine their suitability. This allows for risk free experimentation where stakeholders can try out *what if* scenarios and have the ability to play out the results of interventions over time.

Health care system simulation has been applied to a number of areas (Homer & Hirsch, 2006), including:

- Disease epidemiology: heart disease, diabetes, HIV/AIDS, cervical cancer, chlamydia infection, dengue fever, and drug resistant pneumococcal infections
- Substance abuse epidemiology: heroin addiction, cocaine prevalence, and tobacco reduction
- Patient flows through emergency and health care
- Health care capacity and delivery in areas such as population health planning, dental care, and mental health, as well as how the health system will be able to cope with natural disasters and terrorist incidents
- Interactions between health care and disease epidemiology
- Performance management
- Evaluation of the performance and impact of information and communications technology applications on health care (Anderson, 2002), including Computerised Physician Order Entry systems and electronic prescribing systems designed to reduce medical errors leading at adverse events.

The common characteristics of these areas are that they have features of dynamic complexity and involve large scale or multiple settings in which it would be difficult to conduct traditional experimental methods.

SYSTEM SIMULATION CAN OVERCOME SOME LIMITATIONS OF RANDOMISED CONTROL TRIALS

Randomised control trials (RCT) are often regarded in health care circles as the *gold standard* for evaluating the effectiveness and quality of health interventions (Muir Gray, 1997), with alternative research methods often judged by the degree to which they approximate RCT design.

While being recognised as the best way to evaluate new drug and treatment interventions, many problems which face the health system are complex, dynamic, and cannot be reduced to a single or linear cause and effect relationship. In these situations conducting RCT experiments may not be possible or ideal. When this occurs, system simulation can provide an alternative means of investigation.

For example, RCTs are not practical when the timeframe is too long (Black, 1996). Changes in population health, the management of chronic illness, and planning for future infrastructure will occur over years, decades, or even a lifetime. Aside from the impracticalities of conducting a study over such durations, in many cases decisions can not be put off that long. A computer simulation on the other hand can replicate many aspects of the experiment in a matter of minutes.

The scale of a study may also prohibit the use of RCTs (Black, 1996). This is especially the case in any study that deals with an entire population, such as epidemics or health delivery and capacity. The simulation researcher on the other hand can simulate the entire population.

Trials may be too expensive to conduct, as is the case with major implementations of information and communications technology or building new infrastructure. Simulation experiments overcome the cost barrier as these are tested *in silico* without the need for expensive purchases or capital investment. In a similar vein, this eliminates wasted money should simulation suggest that that a particular piece of new technology or infrastructure is not beneficial.

In certain circumstances an RCT may be unethical or too risky to conduct, especially when it may place the health and welfare of participants or other people at risk. This may arise from inherently risky experiments such as quarantine measures to isolate an outbreak of the H5N1 strain of avian influenza where the consequences of failure would be the deaths of many people, or an unanticipated and unexpected consequence of an experiment which could harm people involved. Again a simulation experiment eliminates the risk to participants and others.

Finally, an RCT is a very controlled and artificial process. It is not always possible to control all perceived and unknown extraneous variables (Black, 1996). Nor should it be desirable to be so controlling, as doing so would ignore that health care is a complex system with high levels of dynamic complexity. In an RCT, isolating the problem may mean removing it from its broader context which does not reflect how a problem behaves in the real world.

In contrast to a trial which attempts to control and limit variance from other sources, a system simulation aims to understand the interactions that produce the observed phenomenon under study. This may involve consideration of various factors that might be considered extraneous or confounding variables in an RCT. The difference is that system simulation focuses on the problem at hand, while RCTs concentrate on the causal relationship between an independent and dependant variable. A further advantage of system simulation is that it allows interventions to be tested prior to implementation, along with numerous other *what if* scenarios and all without additional expense in a risk free environment.

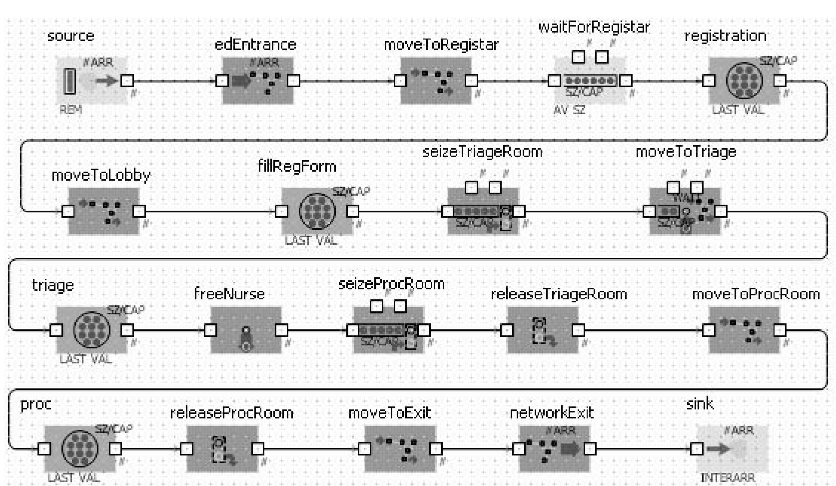
SYSTEM SIMULATION METHODS

Discrete Event Simulation

There are a number of simulation approaches which can be adopted. Discrete event simulation represents a system as a chronological sequence of events. A change of state in the system is marked by events that occur at a set of discrete points in time (Banks, Carson, Nelson, & Nicol, 2004).

Discrete event simulation is commonly associated with cost analysis, queue, inventory, and operation design and management problems. In the health care setting, it has been widely applied to patient scheduling, routing, and admissions, resource and staff allocation, and health services planning (Jun, Jacobson, & R., 1999). Its applications extend beyond the clinical setting and include the management of infection and communicable disease, costs of illness and economic evaluation, and screening (Fone et al., 2003). This can be seen in the assessment and design of health policies such as HIV intervention programs (Rauner, Brailsford, & Flessa, 2005), diabetic retinopathy screening policies (Brailsford, Davies, Canning, & Roderick, 1998), prevention and treatment strategies for coronary

Figure 2. A simple discrete event model of the emergency department in a hospital. Once patients enter the hospital they register and then wait to be seen by the triage nurse. Patients are seen in the order in which they arrive and must wait until a nurse and a triage room are available before being seen.



heart disease (Cooper, Brailsford, Davies, & Raftery, 2006), and resource requirements for renal treatment (Davies, 2006).

Agent Based Simulation

Agent based simulation reveals the global behaviour of a system as structures and patterns emerging as a result of repetitive and competitive local interactions between individuals, or "agents," and their environment (Axelrod & Tesfatsion, 2007; Bonabeau, 2002). The behaviour of each individual is defined by its encoded characteristics and rules for interactions which may be fixed or changed by an interaction with an agent or the external environment (Epstein & Axtell, 1996). This modelling technique also captures how a system learns, adapts, and self-organises itself to maintain order (Begun, Dooley, & Zimmerman, 2003). Agent based modelling is useful for describing systems that are open, complex, and with distributed control and resources. It is predominantly used for modelling epidemics and human social phenomena. Within the health arena, agent based modelling has been used to analyse the management of disease epidemics (Epstein & Axtell, 1996) like HIV epidemics (Teweldemedhin, Marwala, & Mueller, 2004), substance abuse (Perez, Dray, Ritter, Dietze, Moore, & Mazerolle, 2006), and health care and service delivery systems (Kanagarajah, Lindsay, Miller, & Parker, 2006).

System Dynamics

Developed by Jay Forrester at MIT in the 1950s, System Dynamics is a method that enhances understanding of the functioning of complex systems over time and in particularly dynamic complexity, by computing system interaction and taking into account nonlinear feedback (Sterman, 2000). System Dynamics uses a high level of aggregation, continuous flows, and focuses on the pattern of behaviour produced by a system, such as increasing costs, decreasing quality, and stagnating waiting times. System Dynamics modelling has been used to study policy interventions to reduce hospital waiting lists (Gonzalez-Busto & Garcia, 1999) and the care delivery capacity for chronic illness management (Hirsch & Homer, 2004) to name just a few.

Dynamical System Simulation

The precursor to System Dynamics, Dynamical Systems and System Dynamics share a number of similarities. Both model a phenomena whose state changes over time, both have continuous flows, and both do this using models of differential equations and other mathematically specified relationships, where the equations describe behaviours within the system (Willems, 1991).

However there are also some differences. Dynamical Systems simulation is usually limited to simulations

Figure 3. An agent based model, Acute Aged Care Interface (Adaptive Care Systems, 2007). This model tracks elderly people (the agents) in the community, where they stay overnight, and their level of disability and disease.

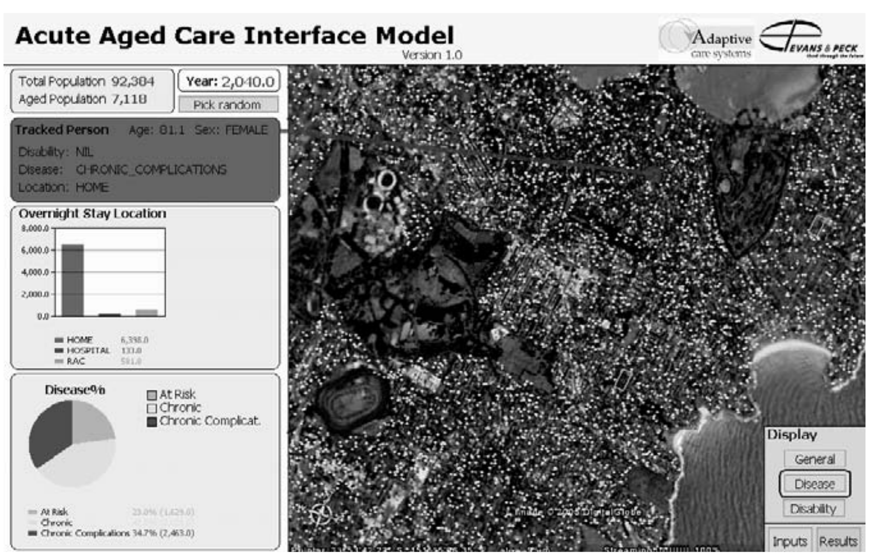
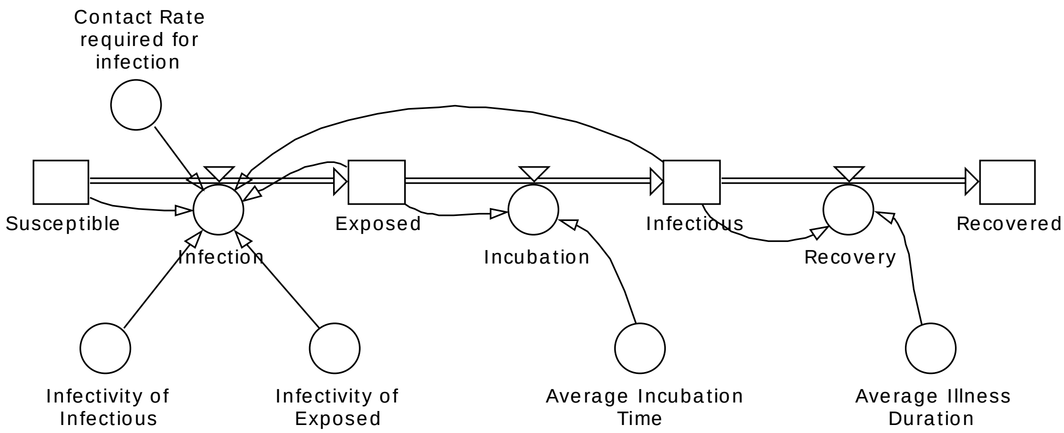


Figure 4. The SEIR model for disease outbreaks represented in a stock and flow diagram used by System Dynamics. The stocks (Susceptible, Exposed, Infectious, and Recovered) represent accumulations of people in each category. The flows (Infection, Incubation, and Recovery) represent the flow of people as they transition from one state to the next. The arrows show feedback, especially how the number of people exposed and infectious impact upon the number of people who become exposed, via infection.



of physical and physiological systems (Borshchev & Filippov, 2004), such as heart beat or respiration. Dynamical Systems simulation is used in the natural sciences, especially physics and mathematics and in physiology modelling in health research.

In contrast, System Dynamics models work at a much higher level of abstraction and while able to model physical phenomena, it is also able to model more abstract policy concepts, such as quality of care and workforce morale.

Multi Method Modelling

Each modelling method is best suited to describe a particular level of abstraction of the problem. System Dynamics outlines the context of the problem making explicit the boundaries of the health system, its structure, and relationships. Discrete event models are process focused and agent based modelling can provide more detail at the individual level as well as a depicting the collective behaviour at the macro level.

Some problems can be elegantly described at one abstraction level by one modelling method however, health care problems span across a wide scale (Bar-Yam, 2006), from the patient, health professionals, and communities to the organisations, institutions, and resources whose primary purpose is to improve peoples' health (WHO, 2006). Multiple levels of abstraction of a health problem, described using multiple modelling paradigms, are required to capture the complexity within and across each level of the health system hierarchy at different time scales.

Multi method modelling, also referred to as hybrid modelling, combines more than one modelling method into one model. Multi method simulation is the execution of this multi method model over time. This process is made simpler with simulation platforms now capable of integrating and cooperating efficiently between different modelling paradigms (Borshchev & Filippov, 2004).

Multi method modelling can be used to develop multiscale simulations that bring together multiple scales of a problem in the one simulation. Multiscale simulation is yet to make its mark in the health industry, however its success in the environmental science discipline and applications to biological systems analysis (Bassingthwaighte, Chizeck, & Atlas, 2006) provide a promising future in health.

CHALLENGES OF SIMULATION

System simulation is not a solution to the problems of health care per se, but instead its value lies in its ability to enhance our understanding of the systemic causes of health system behaviour. This knowledge then becomes the basis for the development of interventions to improve the system, which can be tested on the simulation. The greatest barriers to the success of system simulation stem from barriers to learning.

A commonly used method of teaching the insights gained through system simulation is to allow stakeholders in a problem to “fly” the simulator themselves. They are able to run the simulation, pull the different policy levers, try each *what if* scenario, and observe the results. However, there is the potential danger with simulation gaming that participants will treat it as a video game where the goal is to win or beat the next person’s highest score. This *Video game syndrome* is a barrier to learning when participants instead of learning, play too much, and think too little (Isaacs & Senge, 1992).

Another risk is that participants will learn the insights of the simulation, but will fail to translate these lessons into the real world. Simulation is a distillation of the problem faced without the distractions of the real world. When participants return to their workplace after flying the simulator, it may be easy for them to get caught back up in treating the surface symptoms of the problem, rather than treating its root cause.

Both of these problems can be overcome with good facilitation of the simulation gaming. A good facilitator will draw attention to the specific purposes of the model, provide the conceptual background, and strongly tie the flight simulator to the real world problem. In this way the focus is on the real world problem, not on the simulation as a game or a concept that is removed from the real world.

There is also a risk that simulations and model can be seen as *black boxes* with people unsure of exactly what is occurring inside the computer to produce the results observed. When a simulation is seen as a black box, it is easy for sceptics to dismiss them, weakening the effectiveness of insights gained. This can especially be the case when the results or course of action the process reveals in counter intuitive.

Modellers can overcome the perception of simulations as a black box and improve the acceptance of the simulation by involving the stakeholders in the development of the simulation (Sterman, 2000). Stakeholders should be encouraged to ask questions, suggest changes, and challenge and test model assumptions. These people then become the champions of the simulation and increase the level of acceptance amongst their peers.

A systematic review of system simulation in population health and health care delivery (Fone et al., 2003) concluded that despite the potential of simulation to inform and improve policy development, there is little evidence of the value or impact of such studies and a lack of reporting on implementation. There is a need for more research to develop a larger evidence base for successful system simulation-based interventions in the field of health care.

Complex problems in health will inevitably affect multitudes of people, each of whom will have differing views on the causal relationships and structures which create the problem. If stakeholders in the problem feel that their perspective has not been adequately addressed in the simulation model, their perceived validity of the model will deteriorate, weakening the possibilities for learning and positive change.

SIMULATION VALIDITY

A system simulation is an abstraction of the real world, aiming to capture the relevant essence of the problem. In this way, a simulation model is a hypothesis of the causal relationships and structures that produce the system behaviour of interest. The usefulness of a simulation will depend on how well it represents the real world problem.

Just as a hypothesis can never truly be proven, a model can never be truly validated (Sterman, 2000). Instead there are a number of techniques that can be used to increase confidence in it. One of the most commonly used is behaviour replication, that is, how well does the simulator replicate historical data.

Testing the simulation with extreme values or conditions helps to ensure that the rules on which the model is built are robust and will respond plausibly when evaluating various policies or determine how the system will respond to shocks.

There needs to be an evaluation of the adequacy of the boundary of the simulation to ensure that concepts important to the functioning of the system are endogenous to the model. Failing to include important concepts may mean that important sources of interactions or feedback are being ignored.

Sensitivity analysis, which determines how sensitive the model is to parameter changes, should be performed on parameters where the initial conditions are not certain. If a model is very sensitive to an estimated parameter, then we need to carefully question why this is so.

Simulation models should also be tested to ensure the structure is correct, that it is a consistent representation of the real world problem.

FUTURE TRENDS

There are a number of developments within the system simulation field which are helping to expand its utilisation. It is pleasing to note that the use of systemic approaches for analysing the health system is increasing (Fone et al., 2003). We believe that a more complete view of the complexity of the health system across its varying levels along with multiscale analysis will be best achieved using a combination of modelling paradigms.

Another important trend concerns the incorporation of human factors in system simulation. Health care systems are influenced significantly by the behaviour of the people within them (Brailsford, Sykes, & Harper, 2006). Currently human behaviours are poorly represented or omitted in health system models causing observed behaviour to depart from the behaviour expected from the modelled system. The incorporation of human factors and behaviour in simulation models will be enhanced by the development of agent based modelling (Brailsford & Schmidt, 2003).

Advances in technology have made modelling and simulation a more practical tool for exploring, understanding, and presenting the behaviour of the health system. Innovative graphics used in simulation animations can create an emotionally engaging, immersive learning environment heightening the users' experience and encourages more relevant, applicable and realistic decisions to be made. Moreover, the use of parallel and distributed "grid" computing will allow the execution of large complex models to be completed in less time, assisting more effective and faster decision making. Simplifying the modelling process will increase the popularity of system simulation in the health industry. This can be achieved by formalising a project management process around model building and including the decision makers in the model design process, using simple-to-use modelling tools with engaging animations, increasing re-use and interoperability of simulation models with the development of model building standards, and a health modelling library comprising a collection of reusable health modelling building blocks. Multidisciplinary collaborative communities supported by dynamic software infrastructures for combining simulation with data online are emerging in systems biology (Swertz & Jansen, 2007) and in time they could extend to a range of health system problems.

Systems simulation has the potential to provide more realistic insights into the relationships and interactions between components within the health system and its resultant behaviour. Marrying real world data with health system models opens possibilities to accelerate the understanding of the systems actual patterns in behaviour and thus enable decisions for positive change to be made quicker and be more relevant. For example, calibrating a biological model, such as the Archimedes model, with data obtained from electronic medical records could facilitate the simulation of evaluation studies and clinical trials with outcomes made available remarkably sooner than studies performed in real time (Eddy, 2007).

CONCLUSION

Not only is health care a large, complex, and dynamic system, but it is also a critically important one. It is a system where the price of failure may be counted in lives needlessly lost. Learning by trial and error places too much at risk. RCTs may take too long to produce usable results, fail to consider key factors, cost too much, not cover sufficient scale, or be unethical or otherwise impractical to conduct.

Traditional methods of management seek to understand and solve complex problems by reducing, isolating, and freezing in time. As we have established, in a complex and highly coupled system such as health care, these approaches ignore dynamic complexity and the effects of feedback. There needs to be an understanding of the systemic structures and interactions that give rise to social issues and problems. It is only by addressing the systemic causes that they can be fully addressed.

System simulation is the term we give to a collection of simulation methodologies that focus on the interactions of systems which provide the means for gaining the understanding of system causes. They also provide the ability to test intervention *in silico* before they are implemented in the real world. In an ideal world, all health policy and reform would be extensively tested in computer simulations before any changes were made because in something as important as the public health system the price of policy failure is high.

H

REFERENCES

Adaptive Care Systems. (2007). Public Models. Retrieved February 13, 2008, from http://www.healthsims.com.au/public.htm

Anderson, J., G. (2002). Evaluation in health informatics: Computer simulation. *Computers in Biology and Medicine, 32*(3), 151-164.

Axelrod, R., & Tesfatsion, L. (2007, February 10). *Online guide for newcomers to agent-based modeling in the social sciences.* Retrieved February 13, 2008, from http://www.econ.iastate.edu/tesfatsi/abmread.htm

Banks, J., Carson, J., Nelson, B.L., & Nicol, D. (2004). *Discrete-event system simulation* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Bar-Yam, Y. (2006). Improving the effectiveness of health care and public health: A multi-scale complex systems analysis. *American Journal of Public Health, 96*(3), 459-466.

Bassingthwaighte, J.B., Chizeck, H.J., & Atlas, L.E. (2006). Strategies and tactics in multiscale modeling of cell-to-organ systems. *Proceedings of the IEEE, 94*(4), 819- 831.

Begun, J., Dooley, K., & Zimmerman, B. (2003). Health care organizations as complex adaptive systems. In S.M.M.a.M. Wyttenbach (Ed.), *Advances in health care organization theory* (pp. 253-288). San Fransisco, CA: Jossey-Bass.

Black, N. (1996). Why we need observational studies to evaluate the effectiveness of health care. *British Medical Journal, 312*(7040), 1215-1218.

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. In *Proceedings of the National Academy of Sciences of the United States of America, 99*(3), 7280-7287.

Borshchev, A., & Filippov, A. (2004). *From system dynamics and discrete event to practical agent based modeling: Reasons, techniques, tools*. Paper presented at the 22nd International Conference of the System Dynamics Society. Retrieved February 13, 2008, from http://www.xjtek.com/files/papers/fromsystemdynamics2004.pdf

Brailsford, D., R., Canning, C., & Roderick, P. (1998). Evaluating screening policies for the early detection of retinopathy in patients with non-insulin dependent diabetes. *Health Care Management Science, 1*(2), 115-124.

Brailsford, S.C., & Schmidt, B. (2003). Towards incorporating human behaviour in models of health care systems: An approach using discrete event simulation. *European Journal of Operational Research, 50*(1), 19-31.

Brailsford, S.C., Sykes, J., & Harper, P.R. (2006). *Incorporating human behavior in health care simulation models*. Paper presented at the Winter Simulation Conference, Proceedings of the 37th conference on Winter Simulation.

Cooper, K., Brailsford, S.C., Davies, R., & Raftery, J. (2006). A review of health care models for coronary heart disease interventions. *Health Care Management Science, 9*(4), 311-324.

Davies, R. (2006). *Use of simulation to determine resource requirements for end-stage renal failure*. Paper presented at the Proceedings of the 37th Conference on Winter Simulation Conference.

Eddy, D.M. (2007). Linking electronic medical records to large-scale simulation models: Can we put rapid learning on turbo? *Health Affairs, 26*(2), 125-136.

Epstein, J.M., & Axtell, R. (1996). *Growing artificial societies: Social science from the bottom up*. Cambridge, MA: MIT Press.

Fone, D., Hollinghurst, S., Temple, M., Round, A., Lester, N., Weightman, A., et al. (2003). Systematic review of the use and value of computer simulation modelling in population health and health care delivery. *Journal of Public Health Medicine*, *25*, 325-335.

Forrester, J.W. (1961). *Industrial dynamics*. Cambridge, MA: M.I.T. Press.

Gonzalez-Busto, B., & Garcia, R. (1999). Waiting lists in Spanish public hospitals: A system dynamics approach. *System Dynamics Review, 15*(3), 201-224.

Hirsch, G.B., & Homer, J.B. (2004). *Modeling the dynamics of health care services for improved chronic illness management*. Paper presented at the 22nd International Conference of the System Dynamics Society.

Homer, J.B., & Hirsch, G.B. (2006). System dynamics modeling for public health: Background and opportunities. *American Journal of Public Health, 96*(3), 452-458.

Isaacs, W., & Senge, P. (1992). Overcoming limits to learning in computer-based learning environments. *European Journal of Operational Research, 59*, 183-196.

Jun, J.B., Jacobson, S.H., & R., S.J. (1999). Application of discrete-event simulation in health care clinics: A survey. *The Journal of the Operational Research Society, 50*(2), 109-123.

Kanagarajah, A.K., Lindsay, P., Miller, A., & Parker, D. (2006). *An exploration into the uses of agent-based modeling to improve quality of health care*. Paper presented at the Sixth International Conference on Complex Systems (ICCS2006).

Kuipers, B. (1994). *Qualitative reasoning: Modeling and simulation with incomplete knowledge*. Cambridge, MA: MIT Press.

Muir Gray, J.A. (1997). *Evidence-based healthcare*. Edinburgh: Churchill Livingstone.

Paich, M., & Sterman, J.D. (1993). Boom, bust, and failures to learn in experimental markets. *Management Science, 39*(12), 1439-1458.

Perez, P., Dray, A., Ritter, A., Dietze, P., Moore, T., & Mazerolle, L. (2006). *SimDrug: Exploring the complexity of illicit drug markets*. Paper presented at the Proceedings of the 50th Annual Meeting of the International Society for the Systems Sciences.

Plsek, P. (2003). *Complexity and the adoption of innovation in health care*. Paper presented at the Accelerating Quality Improvement in Health Care: Strategies to Speed the Diffusion of Evidence-Based Innovations.

Plsek, P., & Greenhalgh, T. (2001). Complexity science: The challenge of complexity in health care. *British Medical Journal, 323*(7313), 625-628.

Proctor, P., Reid, W., Compton, D., Grossman, J.H., & Fanjiang, G. (2005). *Building a better delivery system: A new engineering/health care partnership*. Washington DC: National Academies Press.

Rauner, M.S., Brailsford, S.C., & Flessa, S. (2005). Use of discrete-event simulation to evaluate strategies for the prevention of mother-to-child transmission of HIV in developing countries. *Journal of the Operational Research Society, 56*(2), 12.

Sterman, J. (2000). *Business dynamics: Systems thinking and modeling for a complex world*. Boston, MA: Irwin/McGraw-Hill.

Swertz, M.A., & Jansen, R.C. (2007). Beyond standardization: Dynamic software infrastructures for systems biology. *Nature Reviews Genetics, 8*, 235-243.

Teweldemedhin, E., Marwala, T., & Mueller, C. (2004). *Agent-based modelling: A case study in HIV epidemic*. Paper presented at the Proceedings of the Fourth International Conference on Hybrid Intelligent Systems (HIS '04) - Volume 00.

WHO. (2006). *What is a health system?* Retrieved February 13, 2008, from http://www.who.int/health-systems/en/

Willems, J.C. (1991). Paradigms and puzzles in the theory of dynamical systems. *IEEE Transactions on Automatic Control, 36*(3), 259-294.

KEY TERMS

Agent Based Simulation (AB): A method of simulation which focuses on the interaction of individual agents. In Agent Base simulations, agents are encoded with attributes and rules governing how they interact with other agents and what attributes may change as a result of the interaction.

Complex System: A highly coupled system where the outcomes of the system are the result of the interactions that occur between its different components.

Discrete Event Simulation (DES): A method of simulation that treats the operation of a system as a sequence of chronological events, with changes in the state of the system being triggered by a discrete event.

Dynamic Complexity: Occurs when systems are dynamic, tightly coupled, governed by feedback; plays out over time.

Dynamical System Simulation (DSS): A method of simulation that is similar to System Dynamics (SD). However in contrast to SD, Dynamical Systems simulation are usually limited to simulations of physical or physiological systems where the relationships are expressed as mathematical equations.

Health System Simulation: The application of quantitative modelling and computer simulation methods to study the interactions between individuals and/or components of the health care system and how these interactions over time produce the observed behaviour which is produced by the health care system.

Multi-Method Simulation: Asimulation which uses more than one simulation method.

System: A combination of parts which form a whole entity.

System Dynamics (SD): A method of simulation that focuses on system interactions over time, which considering feedback effects. System Dynamics models are made up of a series of differential equations and are continuous. They are highly aggregated, focus on patterns of behaviour, and can incorporate abstract concepts. It is a useful method for policy level analysis.

Healthcare Information Systems and the Semantic Web

David Parry
Auckland University of Technology, New Zealand

INTRODUCTION

The World Wide Web (WWW) is a critical source of information for health care. Because of this, systems for allowing increased efficiency and effectiveness of information retrieval and discovery are critical. Increased intelligence in Web pages will allow information sharing and discovery to become vastly more efficient .The semantic Web is an umbrella term for a series of standards and technologies that will support this development.

BACKGROUND

The Web and Healthcare Information Systems

The early development of the Internet, from ARPA developments and the Internet protocol in the 1960s and 1970s to increasing use of e-mail and text-based systems, led to the development of HTTP and HTML in the 1990s (Barry et al., 1997). The Web has become a critical source of information and communication among people in all domains. The world of medicine on the Web, including such concepts as eHealth, have become areas of great academic interest (Pagliari et al., 2005). Thus, the Web and health care are increasingly co-dependent.

Finding the required information in a search of HTML documents is difficult, as HTML is really a formatting language for humans rather than for indexing. Of course, Web-based systems such as search engines or digital libraries are available, but these rely on indexing either as a set of keywords drawn from a limited vocabulary or from sometimes unreliable parsing of the document. Knowing what information a Web page contains, whether it is a medical record, a clinical result, or an academic paper, is still difficult.

The Semantic Web

Tim Berners-Lee has coined the term "Semantic Web" (Berners-Lee, Hendler, & Lassila, 2001). This term has been enthusiastically adopted and generally includes any work intended to allow the meaning of data contained in Web resources to be made accessible to software agents or humans.

In essence, a semantic Web document is a collection of three parts:

- The document that contains the information, marked up in a suitable way
- A document that contains the rules for interpreting the markup (Namespace)
- A document that links the information with the namespace (i.e., which set of interpretation rules is being used).

There has been a great deal of work under the auspices of W3C to develop standards for the semantic Web, in particular representations using XML. Two of the major efforts are the Web ontology language (OWL) (Smith, Welty, & McGuinness, 2004) and the resource description framework (RDF). In terms of information retrieval, the semantic Web will allow a user to express what they mean to find and find objects that satisfy that request without regard to the language or syntax of the request. The semantic Web implies ways of representing the meaning of documents and constructing queries to discover that meaning. If possible, such a process will be as automatic as possible and incorporate the advances in computing power available to users on the desktop and at the server.

An example may make this system clearer. The example is taken from the W3C example page (http://www.w3.org/2000/10/rdf-tests/)

The original RDF document is:

```
<?xml version="1.0"?>
<rdf:RDF
  xmlns:a="http://description.org/schema/"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:oiled="http://img.cs.man.ac.uk/oil/oiled#"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:daml="http://www.daml.org/2001/03/daml+oil#"
  xml:base="file:/C:/downloads/RDF-Model-Syntax_1.0/ms_4.1_1.rdf">
 <rdf:Statement>
   <rdf:subject rdf:resource="http://www.w3.org/Home/Lassila"/>
 <rdf:predicate rdf:resource="http://description.org/schema/Creator"/>
  <a:attributedTo>Ralph Swick</a:attributedTo>
  <rdf:object>Ora Lassila</rdf:object>
 </rdf:Statement>
</rdf:RDF>
```

This can be displayed as shown in Figure 1.

This can be expressed in words as “Ralf Swick and Ora Lassila have the attributes of creators of RDF.”

The flexibility of this approach can be seen by the fact that the definitions of what RDF is and what “creator” means can be located in other files. Thus, agreed meanings can be used across multiple data files.

Often it may be useful to think of the meanings of items in documents as related to an ontology and that there is a standard for the expression of ontologies on the WWW. This standard—Ontology Web Language (OWL) (Smith, Welty, & McGuinness, 2004)—allows ontology relations to be coded in a machine and human readable form within files that can be located on the WWW.

An example of an OWL statement is shown in Figure 2 (the example is taken from Smith, Welty, & McGuinness, 2004.

This example shows the definition of an equivalent class—items that satisfy the conditions given in the equivalent class construction are regarded as equivalent. In this particular example, things from Texas are assigned to be “Texas things.” This project is still at a fairly early stage but obviously represents a potentially beneficial approach to standardization that may be useful for the deployment of the fuzzy ontology described later.

What both these approaches do is allow the page author to identify to agents or humans the intended meaning of various parts of the document. The RDF form does not need the existence of an underlying ontology, but in many cases, an ontology does exist, which makes consistent namespace creation easier. Of course, systems such as HL7 already have what are effectively namespaces, but the semantic Web approach means that the author provides the data, the possible values of the data, and the potential meaning of the data in a format that allows shared meanings. This

Figure 1. XML document construction

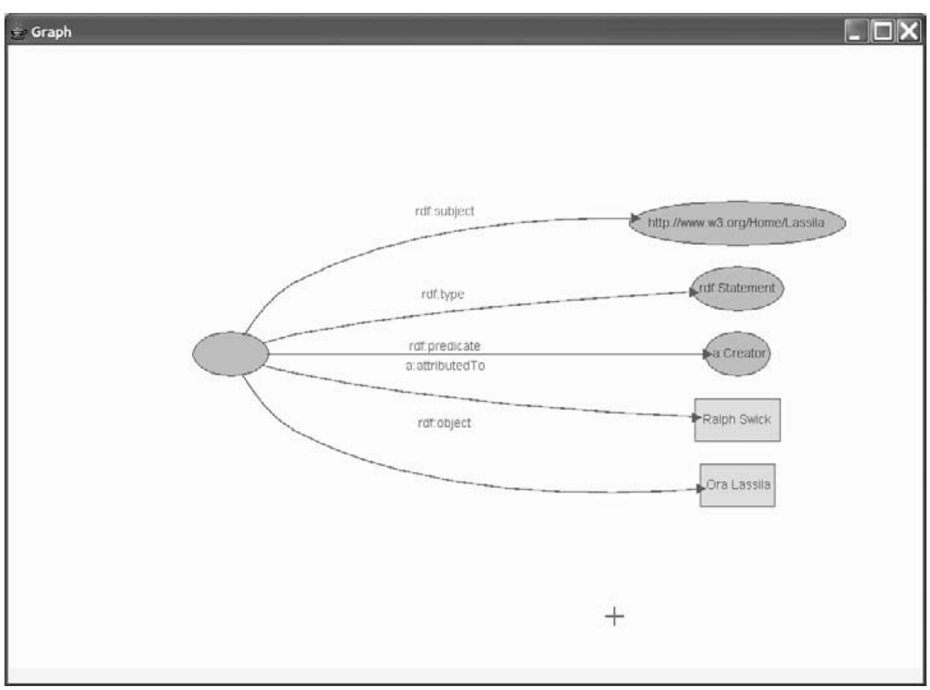


Figure 2. OWL example

```
<owl:Class rdf:ID="TexasThings">
 <owl:equivalentClass>
  <owl:Restriction>
   <owl:onProperty rdf:resource="#locatedIn" />
   <owl:someValuesFrom rdf:resource="#TexasRegion" />
  </owl:Restriction>
 </owl:equivalentClass>
</owl:Class>
```

makes the approach very powerful for semistructured information or approaches where only a subset of the information is needed.

Systems for encoding semantic information do not have to rely on XML or other general-purpose systems. An extremely general-purpose approach to adding semantic information to documents is the WorldNet approach described in Mihalcea and Mihalcea (2001). In WorldNet, there is a number of keywords, each of which has a large potential set of attributes such as synonyms, hyponyms, related terms, and so forth, which are used for searching. Documents have their own schema that includes information about the position of e-words within the document.

HEALTH INFORMATICS AND THE SEMANTIC WEB

The use of semantically meaningful XML structures is not new in medicine. The entrez database that is used to access PubMed has a standard XML output that allows users and agents such as the entrez e-search tools to identify elements of a document record. Similar systems are available for gene and protein libraries.

The use of Semantic Web technologies for a health database is described in AnhÃ¸j, 2003. In many cases, the use of ontologies has predated the implementation of Semantic Web formulation in the medical domain. In the previous terms, the reference to namespace is implicit, as the namespace is determined by the ontology that is being used. This approach is more powerful in some respects, as extremely complex relations and workflow can be encoded into complex ontologies, but less flexible in that all the documents have to refer back to the same namespace.

Web-based ontologies can be seen as a part of implementing the Semantic Web. The Stamford medical informatics group suggest their language—PROTÉGÉ-2000 (Noy, Sintek, Decker, Crubézy, Ferguson & Musen, 2001)—as a means of representing ontologies on the Web. The central role of ontologies in knowledge management is emphasized by Staab and Maedche (2001).

A number of systems have been designed to allow communication between ontologies; for example, XOL, an otology exchange language implemented in XML and developed from Ontolingua (Karp, Chaudhri, & Thomere, 2004). The Ontolingua server (Stanford University Knowledge Systems Laboratory, 2001) allows online collaborative editing of ontologies, and the knowledge system laboratory is also the home of a library of ontologies (e.g., the enterprise ontology and others).

The MeSH system also can be easily represented as an ontology, although in this case an object can be allowed to be a member of more than one leaf element at the lowest level. This is because a particular document can be "about" many different and closely related subjects. It is also possible to have the same leaf element incorporated in a number of different branches. This creates problems for some representation systems, and artificial distinctions have to be made between the same terms, depending on from which branch it has come.

A keyword or index term hierarchy can be seen as a particular class of ontology. At the same time, documents can be analyzed in terms of such ontologies. In this case, the attributes used to define a document are themselves divided by navigating the tree that represents possible classifiers. In the information retrieval case, it is highly likely that documents will possess many attributes, some of which may tend to place it in differing parts of the ontology. For example, a page produced by a sufferer's charity about a disease may have attributes such as brevity, low reading difficulty that indicates it would be located in a tree associated with public information, but it may have valuable links to research teams and means of getting funding that would make it suitable for researchers and located in that tree. An XML result from PubMed can be seen as an example of a Semantic Web document, where the namespace is not directly referenced but can be obtained from the publisher.

```
<?xml version="1.0"?>
<!DOCTYPE PubmedArticleSet PUBLIC "-//NLM//DTD PubMedArticle, 1st January 2006//EN" "http://www.ncbi.nlm.nih.gov/entrez/query/DTD/pubmed_060101.dtd">
<PubmedArticleSet>
<PubmedArticle>
 <MedlineCitation Owner="NLM" Status="MEDLINE">
  <PMID>15204612</PMID>
  <DateCreated>
   <Year>2004</Year>
   <Month>06</Month>
   <Day>18</Day>
```

Not only knowledge sources can be represented in this way. The Electronic Health Record (EHR) can use Semantic Web technologies, for example, by identifying regions that correspond to patient ID, date of birth, or diagnosis.

Current Research Projects in this Area

The National Library of Medicine (http://www.nlm.nih.gov/healthit.html) runs a number of large-scale projects in this area; in particular, the UMLS work mentioned earlier. This forms a basis for a great deal of work. Underlying the concept of a Semantic Web is that of ontology. The protégé research group at Stanford (http://protege.stanford.edu/) is continuing to develop interactive tools and ontology products.

The WWW consortium (W3C) has active programs developing and promoting OWL and RDF (http://www.w3c.org). Research in ontology development includes the use of fuzzy ontologies to better match uncertain classifications (Parry, 2004; Tho, Hui, Fong & Tru Hoang, 2006) and the development of ontologies for specific areas such as the gene ontology machine (http://www.**geneontology**.org). Applications of Semantic Web approaches to health care include drug safety (Stephens, Morales & Quinlan, 2006) and the investigation of e-health quality and reliability (Eysenbach, Lampe, Cross & Brickley, 2000). A particularly strong effort is being made to use semantically enhanced web services in the medical domain in the Artemis project (Asuman et al., 2006).

FUTURE TRENDS

The Bioinformatics community has a number of applications ready for Semantic Web approaches (Wang, Gorlitsky & Almeida, 2005). The wide variety of interfaces to the various databases in this area remains difficult, and a standardized semantic approach seems particularly likely to arrive. The latest release of HL7, version 3 (Dolin et al., 2001) is explicitly based around semantics of documents. Many applications are based around data-sharing via XML documents and Web services, and standards are emerging. Semantic standards—what a particular piece of information means—is intrinsically more complex for more sophisticated information, and it may be that hybrid approaches are used. That is, items such as name, patient number, drug therapy, and so forth are easily standardized, but more abstruse concepts such as prognosis or lesion appearance may not be so easily represented. The Semantic Web architecture will become more attractive for purely internal projects as tools for XML and RDF become more effective.

In business terms, there appears to be a role for brokers and intermediaries between systems. There are potential issues with privacy of medical data in that Semantic Web storage should be more comprehensible than proprietary formats, but health care systems always have had this problem. It seems likely that the open source movement will be particularly adept at adopting this approach, which will lead to increased use of the EHR (Middleton, Hammond, Brennan & Cooper, 2005). In terms of health knowledge and information retrieval, interoperability between systems seems the greatest possible benefit and, in particular, the development of intelligent browsers with agent technology that can vastly increase the power of search by personalizing it and putting it in context.

CONCLUSION

The Semantic Web is based around a number of fairly simple standards, but adoption of those standards in the health care domain is sometimes difficult. It is possible that Web publishers will generally adopt Semantic Web coding standards to allow their documents to be used by increasingly sophisticated agent-based systems. The vision presented in Shadbolt, Hall & Berners-Lee (2006) is attractive and is happening for particular domains. The huge bulk of Web pages that have not been revised for ease of semantic understanding seems to pose a problem. However, HTML coding was seen as difficult in 1996, so as more editors become available and the utility of Semantic Web applications becomes obvious, this may not be a vain hope. Undoubtedly, Semantic Web approaches to information publishing and sharing are likely to be a key aspect of work in health care information in the future.

REFERENCES

AnhÃ¸j, J. (2003). Generic design of Web-based clinical databases. *J Med Internet Res, 5*(4), e27.

Asuman, D., et al. (2006). Artemis: Deploying semantically enriched Web services in the healthcare domain. *Inf. Syst., 31*(4), 321–339.

Barry, M. L., et al. (1997). The past and future history of the Internet. *Communications of the ACM, 40*(2), 102–108.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic Web. *Scientific American* (May 2001), 29–37.

Dolin, R.H., et al. (2001). The HL7 clinical document architecture. *J Am Med Inform Assoc, 8*(6), 552–569.

Eysenbach, G,Y.G., Lampe, K., Cross, P., & Brickley, D. (2000). Quality management, certification and rating of health information on the net with MedCERTAIN: Using a medPICS/RDF/XML metadata structure for implementing eHealth ethics and creating trust globally. *Journal of Medical Internet Research, 2*(suppl2)(e1).

Karp, P., Chaudhri, V., & Thomere, J. (2004). *XOL—Ontology exchange language*. Retrieved September 1, 2002, from http://www.ai.sri.com/pkarp/xol/xol.html

Middleton, B., Hammond, W.E., Brennan, P.F., & Cooper, G.F. (2005). Accelerating U.S. EHR adoption: How to get there from here. Recommendations based on the 2004 ACMI retreat. *J Am Med Inform Assoc, 12*(1), 13–19.

Mihalcea, R.F., & Mihalcea, S.I. (2001). *Word semantics for information retrieval: Moving one step closer to the semantic Web.* Proceedings of the 13th International Conference on Tools with Artificial Intelligence, Dallas, Texas.

Noy, N., Sintek, M., Decker, S., Crubézy, M., Ferguson, R., & Musen, M. (2001). Creating semantic Web contents with Protégé-2000. *IEEE Intelligent Systems* (March/April), 60–71.

Pagliari, C., et al. (2005). What is eHealth (4): A scoping exercise to map the field. *J Med Internet Res, 7*(1), e9.

Parry, D. (2004). Fuzzification of a standard ontology to encourage reuse. In *Proceedings of the IEEE International Conference on Information Reuse and Integration (IEEE IRI-2004)*, Las Vegas, NV.

Shadbolt, N., Hall, W., & Berners-Lee, T. (2006). The semantic Web revisited. *Intelligent Systems, IEEE [see also IEEE Intelligent Systems and Their Applications], 21*(3), 96–101.

Smith, M.K., Welty, C., & McGuinness, D.L. (2004). *OWL Web ontology language guide*. Retrieved March 1, 2004, from http://www.w3.org/TR/2004/REC-owl-guide-20040210/#OntologyMapping

Staab, S., & Maedche, A. (2001). Knowledge portals: Ontologies at work. *AI Magazine, 22*, 63–85.

Stanford University Knowledge Systems Laboratory. (2001). Ontolingua ontology editor. Retrieved January 3, 2004, from http://www-ksl-svc.stanford.edu:5915/&service=frame-editor

Stephens, S., Morales, A., & Quinlan, M. (2006). Applying semantic Web technologies to drug safety determination. *Intelligent Systems, IEEE, 21*(1), 82–88.

Tho, Q.T., Hui, S.C., Fong, A.C.M., & Tru Hoang, C. (2006). Automatic fuzzy ontology generation for semantic Web. *Knowledge and Data Engineering, IEEE Transactions, 18*(6), 842–856.

Wang, X., Gorlitsky, R., & Almeida, J.S. (2005). From XML to RDF: How semantic Web technologies will change the design of 'omic' standards. *Nat Biotech, 23*(9), 1099–1103.

KEY TERMS

Electronic Health Record: A system of storing information about a patient to assist in diagnosis and treatment.

HL7: A medical message system standard.

MeSH: Medical Subject Headings. Index terms used by NLM.

Ontology: A rigorous and exhaustive organization of some knowledge domain that is usually hierarchical and contains all the relevant entities and their relations.

Semantic: Of or relating to meaning or the study of meaning.

PubMed: Web portal to the National Library of Medicine Bibliographic Database.

W3C: The World Wide Web steering committee.

WWW: The World Wide Web. A system using data and program protocols that allow simple publishing and transfer of information.

XML: Extensible Markup Language. A standard set by W3C for documents that can include description of the meanings of items.

Healthcare Training for the Greek Municipalities' Citizens in the Program of "Home Help" with the Medical Educational Web Sites

Bill Ag Drougas
Higher Technological Institute of Epirus, Greece

INTRODUCTION

Today, the Internet is one of the most useful tools for information, education, business, and entertainment. It is one of the modern technology tools giving us many applications. One of the most important applications of the Internet is the health educational Web sites. These sites can be used not only by researchers but also by everyone easily taking important information for continuing self-training. In the division of public health, this part is very important. Medical Web sites give us access to information about our health or about prevention methods. In Greece today, there is a new 12-hour program offering home health help and care to citizens. This program is under the organization of the local municipalities and the affiliation of the Ministry of Health.

BACKGROUND

In Greece and especially in the agricultural municipalities, there are many old people who must have healthcare at their home. The Greek government established a serious program for public health. But still this program, after some years since the beginning of its application, continues with many problems. New technologies and the Internet may help local municipalities and their citizens.

The Home Help Care program was established by the ministry of Health in Greece to help people in Greece in their homes. For the application, there were about four to eight people working in the program with various levels of training but under the same direction in order to help people in their homes, giving them health and psychological help.

The administration, application, and recognition of the program Home Help is under the recognition of the Administration Work Group (AWG), which established it with the No.DY1d/oik. 10024/1.6.2001 recognition No. of the Government (FEK 726/τ.B/2001 paragraph 18 article 18 N. 2503/1997 (Government Newspaper FEK 107/rA/97) and still exists today (Moraiti, 2005).

This program offers a first-step healthcare to citizens, so it is important to help them use the contemporary technologies for self-training or for simple self-help and protection.

ABOUT THE RESEARCH

There is an enormous number of programs today in the Greek Kapodestrian Municipalities. One very important program with many positive applications is the Home Help Care program. Many individuals can help themselves along with the affiliation and the help of Home Help Care program employees in the municipality. This research recognized the problems from the application of this program.

This research took part in western Greece. The employees of the program asked to help giving the questions to the citizens and taking back the answers. Then all the answers were organized carefully. There were many results with an important statistical study and many different diagrams. We were able to understand what happens with the program of Home Help Care and what we must do in order for this program to work more efficiently. In this research, there were many statistical tables and many more statistical diagrams, but we will present only a few of them.

ISSUES

The existing Greek geographical status today makes more difficult the application of healthcare, especially at the municipalities far away from the big cities or to the villages where there are many climatic problems that continue for many months per year.

The traditional way to offer help to people in various places makes the application of this program very difficult. By using the periodical visits in every home and spending much time for this either in transportation or staying in the same house for a long time, the program cannot work fast for a simple or combined health problem. People working in the program today are not able to visit all the persons once or more times per week because the number of them is very big and the claim for psychological help is bigger every day with various problems, especially with older persons. Sometimes they must work with the same person more than one time and for many hours.

The effort of employees at time management gives them the probability for more dangers driving many kilometers to every village far from the center, which happens every day.

There are only a few persons working in the program, less 4 and maximum 7 persons, offering home help and care. So it is very difficult for them to work every day personally with every citizen (Moraiti, 2005) They are not able to visit all the persons in their home in different villages many kilometres away. Another problem is that the official number of inducted persons in the program asking for home help is very big in every municipality. This problem is bigger if someone works with people with different problems and more than three or four times per person. Especially if the people are over 60 years old, this is more difficult.

Sometimes they must visit and work with the same person for more than one time per week. This is what gives us the opportunity to claim that the IT technologies will help the program to continue with success. On the other hand, municipalities must work hard to induct new technologies and continue to the modern philosophy with many vital technical changes (Drougas, 2005b). Some of these changes are summarized in Table 1.

A network system will help people have access with the control center, even if they are many kilometers away from the center (Drougas, 2004). After searching the selected answers from the questions given to the people associated with the program, we can say that the most serious problems are summarized in Table 2.

All these results from the research give us the icon that the program today continues very slowly. Internet technologies and a self-training methodology program will help both employees and citizens in the next years with success. This will be a part of the applications of the contemporary network systems WAN and LAN, which will help citizens have local or national and

Table 1. Summary of the technical changes in telecommunications in the municipality

- New connections in affiliation with the local telecom organizations or companies
- ISDN Networks
- LANs Local Area Networks
- WANs Wide Area Networks
- MANs Metropolitan Area Networks
- Internet TCP/IP
- Intranet communications
- Extranet communications
- VPN Virtual Private Networks

Table 2. Summary of the vital problems from the application

- The low level of the citizens education
- The big number of the new Kapodistrian municipalities with many villages many kilometers far
- The small number of the employees in the program
- The big number of the citizens served in the program
- The personal visit in every person
- The difference of the health problems
- Geographical problems
- More persons are older than 60 years old
- Telecommunications problems
- Economical and other time problems

international access to Internet information using their home computers (Apostolakis, 2002).

It is time to train all citizens, the municipality's administration office employees, and the Ministry administration office employees in IT philosophy of network for the e-health education and applications (Drougas, 2005b).

Citizens are not able to use the Internet connections, and many of them are unable to search and use or understand the Internet medical information Web sites. Also, we must say that there are only a few Internet sites in the Greek language today with serious information about healthcare.

This study explains why the case of Greek agricultural municipalities are different than other central and how they can help families understand the use of the Internet medical resources for a self-training, rehabilitation, first aid, family health, and so forth.

The low level of education in the villages makes the work of the employees more difficult because sometimes there is much time to spend with only one person to help him or her for one or more problems. This is a phenomenon appearing in the agricultural areas in the same municipality today. Telecare and telecommunications can send information many kilometers from the control center with best quality, security, and in less time. We are able to repeat this many times per day or per week and for different places in the municipality and to different persons.

The following statistical analysis gives us the icon about the 300 citizens in the research. Searching the numbers in the following presented tables, it is easy

Table 3. Males and females

TABLE 3		Less than 450€		Over 450 €		Total	
Sex	Male	62	79,49%	16	20,51%	78	100%
	Female	198	89,19%	24	10,81%	222	100%
	Total	260	86,67%	40	13,33%	300	100%

Table 4. Citizens' health level

TABLE 4

HEALTH	No	%	%
Bad	48	16,00	16,00
Middle	201	67,00	83,00
Good	51	17,00	100,00
Total	300	100,00	

Health Problems	Persons	%
Kinetic Problems	60	20,00
Heart Problems	109	36,33
Psychosomatic	26	8,67
Lung Problems	16	5,33
Neurological	33	11,00
Eyes Problems	5	1,67
Rheumatics	6	2,00
Gland Problem	17	5,67
Cancer	18	6,00
Allergy	4	1,33
Other	6	2,00
TOTAL	300	100,00

Table 5. Personal economic status

TABLE 5		MORE HEALTH CARE FROM THE PROGRAM					
PERSONAL ECONOMICAL STATUS		Much – Very much Care		A little more care		TOTAL	
	BAD	18	11,32%	141	88,68%	159	100%
	MIDDLE	4	3,64%	106	96,36%	110	100%
	GOOD	0	0%	31	100%	31	100%

Table 6. General information on program's satisfaction

TABLE 6 General Information	No. persons	Total	Part
For the Program		%	%
Less satisfaction	9,00	3,00	3,00
Much Satisfaction	152,00	50,67	50,67
Very Much Satisfaction	139,00	46,33	46,33
TOTAL	300,00	100,00	100,00

for someone to understand that the employees in the program must work very hard to serve all the people. The number of male and female persons in the program in our research are presented in Table 3. Forty people did not answer.

Their personal health problems are summarized in Table 4.

Table 5 includes data about the personal health and the economical level.

Table 6 presents general information about the satisfaction from the program's application with the divisions of less, much, and very much.

INTERNET SITES ENVIRONMENT IN THE PROGRAM

The program does not associate with any Internet connection and serves people only with personal visits. In the central office, there is not any technician or anyone working in the IT technologies professionally. The program does not use any database, and there is not any future plan to train citizens for a family or personal care via IT and other technologies (Sasiakos, Papastefanatos, & Lazakidou, 2006).

Personal visits are really a big problem in the program where simple problems can change everything very quickly. Using the experience from the past inductees in the program and the experience of the employees, establishing a database may induct always new data, and an effective program like this will continue as a contemporary telehome care program.

With this method of distance advice or with distance help applications, many people will have the opportunity to contact the database using information from this base together with other citizens at the same time and from different places in the municipality. After this, anyone may have answers for his problem directly using the electronic "Tele-Municipality Advisor" for any simple or combined problems. This will help anyone when the visit is not necessary to them personally spending much time.

USING THE HEALTH SITES

Searching the Internet sites, we are now able to say that all the Web sites that have been published by Academic Institutions and professional organizations were of higher quality than those published by any private practice groups or any profit organization, although they gave us much important information.

In Figure 1 the statistical results are shown for how the computer and the Internet were used for the health

Figure 1. The use of contemporary technologies from doctors nurses and hospitals

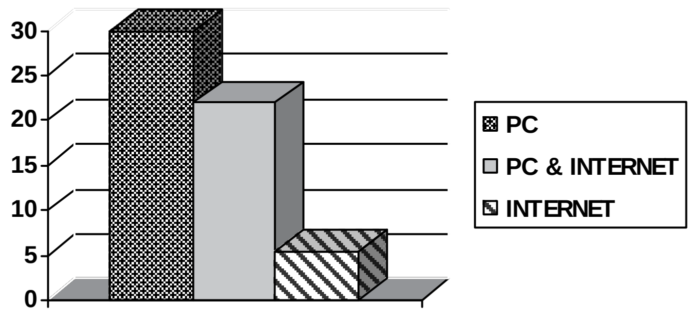
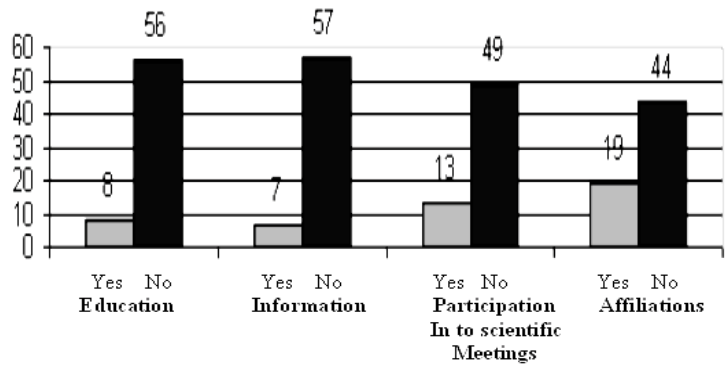


Figure 2. Types of Internet search by the health professionals



help from doctors, nurses, hospitals, and medical companies. 5.5% use the Internet, 30% use only a computer, 22% use both the computer and the Internet.

This statistical diagram shows the answers from health professionals of how the traditional methodologies help them in their work.

All this is recognized from an important research in 2003-2004 for the health professionals to recognize the existing problems in their work area the technological problems and the upcoming induction of the Telematics in medicine applications**.**

Searching the various problems of the affiliated people with the program (HHC), we found that many people today in Greece between 40 and 70 years old are not able to use the modern technologies, especially if they do not have a diploma. Also, many of them do not have a personal computer. There are also many problems with the connecting lines in Greece, and the Internet is still very expensive.

Serious problem were found when we asked them to read and understand information presented on important medical sites. Many of them were not able to organize this information. We must say that health information presented from various sites would be better to follow a video and graphics for self training, audio, or other video files for better education. This will help the old people in the program to better understand the practical application and also people with special needs.

AN IT MODEL FOR THE HHC PROGRAM

In the direction to help people read and train using correct medical information, the Ministry of Health in Greece must create a central database with simple and combined information in the Greek language selected from health professionals and medical bases or scientific magazines and laboratories.

Let's see the model for an Internet self-training from Web pages in the Home Help Care program. This model may help both employees and the people who are associated in the program.

The model is summarized in the next eight points:

- Scientific and medical Web pages selected from the Ministry of Health office
- A new central database will enclose all the medical information with access to other Web pages
- The Ministry of Health will establish a local administration office for the program in every municipality with a central administration
- The local office will have to refresh periodically the local database LDB with access to the central database CDB
- The local office will associate with the program's employees to give and select information about the application of the program
- Citizens who will be in affiliation with the program will have access to the central office and to the local database with a personal electronic address
- Citizens will have the opportunity for self-training and education in different levels and in different subjects periodically
- Citizens will have a dialogue with other persons from the program and scientists and participate in online seminars and information lessons

Table 7 summarizes the answers about IT.

Figure 3 presents the icon from the answers of Table 7.

There is an important machine today in the United States. This is a platform of double direction communication for the healthcare problems. The name of this is Motiva, which works between the TV canals. This program can help people in their homes as a virtual doctor for many health problems, giving information about different health problems. Motiva offers today the following applications:

- Information about one health problem and the vital steps for self-information.
- Personal help from the doctor and communication with the other family members.
- Reminder for medicines and care for personal health.

Table 7. Answers about the IT use

Question	Yes
Do not like computer technologies	20%
You know that it is a useful tool	62%
Do you have a PC in your home?	4%
You think that the HHP uses a computer	8%
Have you heard of the Internet?	70%
You like to ask a computer for your problem	45%
You like to ask a doctor something online	80%
You want to learn to use a computer	54%

HOME HEALTHCARE COMPANIES IN GREECE

Some of the most known companies in Greece offering medical machines and programs in the healthcare:

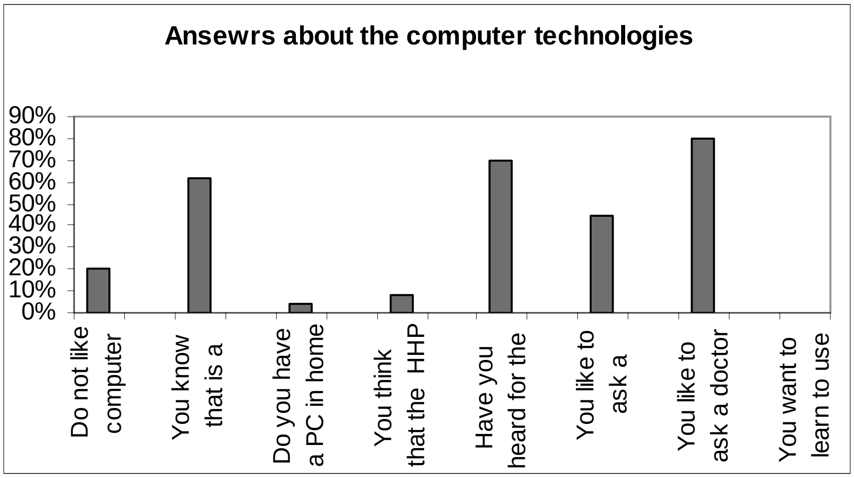
- Advance Products
- Akoustica Medica
- AlfaTech
- Anthos Hellas
- Bio-Spectrum
- BioChem Diagnostics
- Celestar
- Cip Hellas
- Ebedent A.E.
- Sigma Medical
- Sterile Hellas
- Ratio
- Nova Medica
- Nivaco

CONCLUSION

The official recognized program with the name Home Help Care from the Greek Health Ministry is a very important program helping people in every place in Greece. People can use very easily the applications offered from the program, but still there are many problems with the big number of the affiliated persons and the small number of the scientists and the employees who work in the program. The induction of the modern Information technologies will change this program in Greece and make it more efficient.

A personal training for everyone as a personal health information training program will give the information by distance, as personal help from the doctor and com-

Figure 3. Answers about the computer technologies



munication with the other family members or a reminder for medicines and care for the personal protection with the personal doctor.

Using the affiliations of the local companies and the administration of the central municipalities' offices, it is possible to induct new telecommunications technologies into the program's applications. IT technologies will help the program to continue with success and to save money and time. Employees and citizens must train in healthcare and protection electronic technologies and machines, which will help both to associate with success.

REFERENCES

Apostolakis, I. (2002). Medical information systems (pp. 79–82, 318–324). Papazisis Publications.

Demiris, G. (2004). Interdisciplinary innovations in biomedical and health an health informatics graduate education. In *Proceedings of The Missouri experience, health and medical informatics applications educational aspects*. Kapodestrian University of Athens.

Drougas, A.B. (2005a). Home care and modern technologies. D-Publ, 25–50.

Drougas, A.B. (2005b). *Tele-care in the program of home help of the Greek municipalities*. Proceedings of the Health and Medical Informatics Applications International Congress, Athens, 110–112.

Lazakidou, A., & Ilioudi, C. (2006). Security in health information systems. In A. Lazakidou (Ed.). *Informatics in Healthcare and Biomedicine* (pp. 9-15). Hershey, PA: Idea Group Reference.

Mantas, J., & Hasman, A. (2002). *Textbook in health informatics*. Athens: IOS Press

Moraiti, A. (2005). Master thesis in the Greek open university, 70–85.

Paraki, G. J. (2006). Knowledge Management in Telemedicine. In A. Lazakidou (Ed.), *Informatics in Health Care and Biomedicine* (pp. 210-219). Hershey, PA: Idea Group Reference.

Sasiakos, K, Papastefanatos, S., & Lazakidou, A. (2006). The user agent architecture and e-learning. In *Health Care and Social Care Handbook of Research on Informatics in Health Care and Biomedicine* (pp. 105–111).

Sasiakos, K., Papastefanatos, S., & Lazakidou, A. (2006). The user agent architecture and e-learning in healthcare and social care. In A. Lazakidou (Ed.), *Informatics in Health Care and Biomedicine* (pp. 105-111). Hershey, PA: Idea Group Reference.

Zvanut, B. (2006). Motivating healthcare students in using ICTs. In A. Lazakidou (Ed.), *Informatics in Health Care and Biomedicine* (pp. 99-104). Hershey, PA: Idea Group Reference.

H

WEB SITES

www.datamed.gr
www.apollo.gr/dev/index.asp
www.cc.uoa.gr/health
asclepieion.mpl.uoa.gr
www.ics.forth.gr
www.ypyp.gr
www.ehto.org
www.nh.gr/chin/tele/index.html
www.iatriki.gr/
www.ygeia.com.cy/
www.ypai.aegean.gr/asterias
www.who.org
www.infosociety.gr
www.ehto.org
www.ics.forth.gr

KEY TERMS

Healthcare: Care, services, or supplies related to the health of an individual. Healthcare includes but is not limited to the following: preventive, diagnostic, therapeutic, rehabilitative, maintenance, or palliative care; counselling, service, and assessment, as well as the sale or dispensing of a drug, device, equipment, or any other item in accordance with a prescription. (Lazakidou & Ilioudi, 2006)

Home Care Technologies: Any technology used to implement home-care telemedicine services; also any device or any instrument for patient monitoring, therapy, or environmental control (Mantas & Hasman 2002).

Home Help Care Program (HHC): The Home Help Care program established from the Ministry of Health in Greece to help people in Greece in their homes. The administration, application, and recognition of the program Home Help is under the recognition of the Administration Work Group (AWG), which is established with the No.DY1d/oik. 10024/1.6.2001 recognition No. of the Government (FEK 726/τ.B/2001 paragraph 18 article 18 N. 2503/1997 (Government Newspaper FEK 107/rA/97), and still exists today.

Kapodestrian Municipalities: The modern municipalities in Greece with the accreditation of the Greek government. The new local municipalities inducted many small or bigger villages under the same administration.

Network: Defined as a set of nodes and connecting lines to describe intricate structures (e.g., neuronal network, computer network, data network (Paraki, 2006).

Telehealth: The removal of time and distance barriers for delivery of healthcare services or related healthcare activities (American Nurses Association). The methodology of serving people by distance to various heath problems and care. This includes health information programs, self or group training, tele-education, and so forth.

Telehomecare: Uses telecommunication and videoconferencing technologies to enable a health-care provider at a clinical site to communicate with patients in their homes (Demiris, 2004).

WAN: Wide Area Network system of communications by information technologies.

World Wide Web: A system of Internet servers that support specially formatted documents in a language called HTML (Hyper Text Markup Language), which supports links to other documents as well as graphics, audio, and video files (Zvanut, 2006).

Heterogeneous Quality Information in Healthcare Marketplace

Kimberlee Schumacher
Northeastern University, USA

Yang W. Lee
Northeastern University, USA

INTRODUCTION

Contemporary healthcare systems are utilizing new information technology to digitize and share previously paper-based information among the legitimate but distributed participants in the healthcare marketplace. A critical stumbling block the participants face in this effort is the questionable quality of the information. Demands for quality information in the healthcare arena are dramatically increasing as information is being used as a basis for evidence-based medical care. Quality information is also critical for management decision-making and benefit design, but more importantly for providing decision choices for healthcare consumers: patients. Although everyone agrees on the critical need for quality information, there is a lack of agreement on what constitutes quality information for healthcare systems.

In this research, we use the classical stakeholder perspective (Butterfield, 2004; Freeman & Reed, 1983) and information quality research (Lee, Pipino, Funk & Wang, 2006) to examine what constitutes quality information among directly participating stakeholders in the healthcare marketplace. We argue that the role a stakeholder plays in the healthcare marketplace affects the nature of quality information each provides and uses, thus shaping the principles that dictate each stakeholder's view of quality information. Our findings, based on participatory observation and the analysis of research publications, provide a basis for a shared understanding of quality information in the healthcare market place.

BACKGROUND

The healthcare industry is at a crossroads. With healthcare costs continuing a trend of annual double-digit increases, employers are in the unenviable position of either absorbing these costs, passing them on to their employees, or dropping insurance all together. As a result, consumers are responsible for more of their overall healthcare costs and are challenging the healthcare system to become more consumer-driven. A virtual healthcare marketplace is developing in which consumers "shop" for their healthcare. As with any marketplace, consumers are becoming cost-conscious and therefore demand transparency of cost and quality information throughout the healthcare system. In this marketplace, five major groups—consumers (patients), physicians, government, healthcare plans, and employers—form the key stakeholders who play a critical role in shaping what constitutes quality information.

Issues and factors associated with technology acceptance, diffusion, and training (e.g., effective use of new systems by doctors and nurses) have been studied thoroughly (Aas, 2001; Bashshur, Reardon & Shannon, 2001; Cloutier, Godin, Gagne & Reiharz, 2005; Hu, Chau, Liu & Tam, 1999; Jayasuriya, 1998; Menon, Lee & Eldenburg, 2006; Raitoharju & Laine, 2006); the information shared, the contents of the systems, has not been studied adequately.

Guided by the stakeholder perspective (Butterfield, 2004; Freeman & Reed, 1983) and research on information quality (Lee, 2004; Strong, Lee & Wang, 1997; Wang, Lee, Pipino & Strong, 1998), this research asks the following question: What constitutes quality information by multiple stakeholders in the healthcare marketplace? This work further examines principles held by each stakeholder, which serve as the basis for different views on quality information.

This chapter contributes to the research in the healthcare arena for both research and practice. For research, the study will expand and deepen the research in healthcare systems by including the analysis on information that healthcare systems design and use. For practice, the research helps to design and use healthcare information considering all stakeholders' perspectives for an effective evidence-based practice of healthcare. Furthermore, the findings of this study will elevate the discussion on establishing measurements and a common definition of quality information for healthcare practice based on the goal of sharing useful and useable information.

QUALITY INFORMATION IN THE HEALTHCARE MARKET

Many recognize the importance of quality information through a crisis. For example, a CEO of a hospital received a letter with a fine for poor-quality information submitted to the state government (Davidson, Lee & Wang, 2004). An insurance provider paid mistaken claims due to its information quality problems, which resulted from its misaligned business and information processes (Katz-Hass & Lee, 2005). A hospital could not make a critical business decision on medical project initiatives due to inconsistent information (Lee & Pipino, 2004). Physicians resisted using the information from the hospital's data warehouse due to the inconsistent and not-believable service records shown on the report (Lee & Pipino, 2004). Worse yet, ad hoc reports on the news evidenced disasters from mistaken surgeries based on poor-quality information.

A crisis, errors, mishaps, financial loss, and an inability to make decisions are often the results of many cumulated and interdependent factors. A key common factor involved, however, is poor-quality information. Inconsistent data fed from various sources are aggregated to produce a report. Financial and medical information is often aggregated to support decision-making in medical care, medical service project prioritization, and physician compensation. Data collectors, those database professionals who store and maintain the data, and data consumers who use this data and further produce data for various medical and business purposes are all involved in designing, collecting, storing, maintaining, distributing, analyzing, and utilizing the same data. Therefore, various aspects of information quality have different impacts along the long value chain of information.

So what is "quality" information as it pertains to healthcare? Among the various dimensions of information quality (Wang & Strong, 1996)—free of error, timeliness, consistency, conciseness, privacy, accessibility, believability, reputation, ease of use, flexibility, security, value added, objectivity, amount of information—what the healthcare market is currently grappling with is accessibility of information. For example, consumers want to know which physicians and hospitals are the "best." The trouble is, who defines what is the best? Currently, the healthcare industry does not have a unified answer to this question. For example, ask a physician, "Who would you go to if you needed heart surgery?" The physicians would tell you that he or she would ask friends and colleagues to find out who the best heart surgeon is. Health plan providers would resort to statistical information that has measures of evidence of a good practice. Regardless of the fact that one answer may be better than others for a particular purpose, shouldn't this information be available to consumers, the patients? Until recently, this information has been placed in a virtual "black box" unavailable and inaccessible to consumers. Magazines such as *U.S. News and World Report* and *Boston Magazine* publish annual lists of the "best doctors," but critics will tell you that those lists are based on reputation and not actual evidential data, and often the doctors and surgeons on the lists are affiliated with the most well known and largest teaching institutions, not rated according to well-articulated objectives and measured outcomes.

The common definition of quality care needs to be acceptable to all stakeholders, and it needs to be practical enough in order to be usable and useful. Currently, there is a race to define healthcare quality and to provide the information. The quality information and the packaging of this information will be a key differentiator in the coming months and years, particularly for health plan providers.

The pivotal event for quality information in the healthcare arena goes back to 1999. In 1999, the Institute of Medicine (IOM) reported that as many as 98,000 people die annually as a result of medical errors, and they called for a national effort to make healthcare safe. The IOM's report, *To Err is Human: Building a Safer Health System* (Kohn, Corrigan & Donaldson, 1999), galvanized a dramatically expanded level of conversation and concern about patient injuries in healthcare, both in the United States and abroad (Leape & Berwick, 2005). The vision required by the IOM charged the Agency for Healthcare Research and Quality to bring together all stakeholders, including payers, to agree on a set of explicit goals for patient safety to be reached by 2010.

Evidence-based medicine is dependent on the collection, design, and use of quality information that is collectively agreed upon by key stakeholders. Emphasis is on the dissemination and accessibility of information so that the evidence can represent and reach clinical practice and thus service patients. Evidence-based medicine is also based on guidelines of how care should be delivered. The evidence is represented in information that reinforces healthcare guidelines and informs the healthcare practice and services delivered.

STAKEHOLDERS IN THE HEALTHCARE MARKETPLACE

The key stakeholders who are critical for defining quality healthcare information are physicians and care providers such as hospitals, health plans, employers, consumers, and the government. Now we turn to identifying each stakeholder and its underlying principles.

Physicians

Principles held by physicians:

- Physicians know what is best for their patients
- Health plans should be hassle free
- Financial compensation should be fair
- Physicians should be involved in defining quality information
- Evaluative information should be based on physician group level instead of at the individual physician level

In general, physicians are concerned about how information is collected and used and by whom, since the disseminated information can have an impact on their own reputation and future service and work. It is critical that information is consistent with what they perceive as good quality. Quality information for physicians, therefore, means that the use of collected information is well understood and known to the physicians in advance and that the information is collected from reputable sources.

A recent analysis on medical practice (Wennberg, 2006) identified that there is widespread variation in the ways physicians practice, which results in widespread variation in both outcomes and cost. Rochester Independent Physician Association (IPA), an organization of 3,000 clinicians including 900 primary care physicians, has developed a set of measures that they use to compare specialists within each specialty. The purpose is both to develop best practices that all of the specialists should follow by learning from top performers, and to work with the bottom outliers. We show the example of their measures, used for quality improvement purposes, below:

- ***Underused measures:*** Measures that the test or procedure is not being performed as often as it should be.
    - **Diabetes testing:** Did the physician perform glucose and cholesterol testing on diabetic patients?
    - **Colon cancer and breast cancer screenings:** Did the physician order or perform the appropriate cancer screening procedure based on the patient's age and gender?
- ***Misuse measures:*** Measures where inappropriate care is given.
    - Antibiotics for viral infections
- ***Overuse measures:*** Unnecessary care.
    - Antibiotics for sinusitis

The Rochester IPA systematically tries to identify overuse and misuse by reviewing treatments for conditions and analyzing local and regional variation by specialty. For example, the IPA found that carpal tunnel syndrome was treated by some physicians as an inpatient procedure and by other physicians as an outpatient procedure. The outcomes were the same, but it cost a lot more to have an inpatient procedure, so the IPA decided it should be treated as an outpatient procedure moving forward.

Evidence-based medicine is dependent on the use of randomized controlled trials and systematic review based on the collected information. Its emphasis is the dissemination of information so the evidence can influence clinical practice.

Evidence-based medicine is also based on guidelines of how healthcare should be delivered. The evidence is represented in information that reinforces healthcare guidelines and informs the healthcare practice delivered. For example, Bruce Bagley, M.D., Medical Director for Quality Improvement at the American Academy of Family Physicians, states that physicians must move from pedigree to performance and that medical board certification is necessary but not sufficient criteria (World Congress Conference, 2006). However, there is widespread recognition that although leading physicians and academics embrace the notion that quality metrics are important to improving quality care, physicians in the trenches have not all come to the same conclusion.

B. Dale Magee, President Elect of the Massachusetts Medical Society, urges the development of a national measurement set so that physicians are not measured according to different sets of measures that each health plan develops (World Congress Conference, 2006). In the interim, he recommends measurement information at the physician group level instead of individual physician level so the data are statistically more accurate. A concern is that health plans sharing this information with members will result in patients deciding who good doctors are based on aggregated and averaged information, despite the fact that the patient is seen by an individual doctor.

Health Plans

Healthcare plan providers want to achieve effective and profitable management of healthcare with minimum cost. Healthcare premiums continue to increase in double digits, which employers and employees cannot continue to absorb. In response to this difficulty, health plans are trying to develop alternative models. For example, some plans are focusing on paying for quality care that in the long run will keep costs down. The healthcare plan therefore needs to determine what quality care is. The Institute of Medicine Report (Kohn et. al., 1999) offers a working definition: "Quality care refers to the care that is based on evidence-based guidelines and seeks to address overuse, misuse and underuse of services."

Health plans have designed incentive mechanisms for improving the quality of care by physicians. The most common mechanisms are pay- for-performance programs, product design, and public reporting. Health plans have been experimenting with various ways to impact provider reimbursement in ways to promote quality care. The most common method is through the "pay-for-performance" program, which resembles a bonus system. If providers score well on quality metrics specified by the health plan, then they are eligible for an annual bonus that may equate to 2% to 15% of a provider's overall reimbursement. These pay-for- performance plans apply mostly to primary care physicians and hospitals and, in some cases, specialty care physicians.

Due to escalating healthcare costs, employers are looking for ways to pass on more of the cost to their employees. This has resulted in the term "consumer directed health plans." Now burdened with increasing financial responsibility, employees (the direct consumers of healthcare) are more involved in their healthcare decisions. Consequently, consumers need more information, and health plans need to provide this information to consumers, including who the highest quality physicians are and what the costs variance is across various physicians and hospitals.

Beginning in 2004, many of the larger health plans, including United, Aetna, Cigna, and some of the Blues, began to release performance information about physician groups and hospitals to their consumers. For example, these national plans worked with vendors, including Subimo and HealthShare Technologies (part of WebMD), to allow their members access to the software products that let consumers identify where they live and compare how hospitals stack up on certain common procedures so that consumers can compare rates for mortality, complications, average length of stay, and in some cases one to three dollar signs to indicate how costly the hospital is.

Principles held by health plans:

- Reduce costs
- Improve quality
- Get and retain profitable membership
- Keep healthcare providers satisfied enough that they remain in the Plan's network

QUALITY INFORMATION AS DEFINED BY HEALTH PLANS

Information on Efficiency

Efficiency can be measured as a per member per month cost based on care a physician or hospital provides, or by bundled costs of treatment categories. High-cost radiology and lab costs are also measured and closely monitored. Many plans work with a company called Symmetry to measure efficiency of providers through Episode Treatment Groups (ETGs) or aspects of a condition. As consumers share more of the financial burden, we may see a rise in their wanting to know who the most efficient providers are and a movement to those physicians. However, there is still a questionable perception that highest cost providers may be the best cost providers: if you pay more, you get more.

Information on Process

Most physician measures are process measures that are examples of medical underuse. Most plans utilize the national measurement set defined by the National Committee for Quality Assurance (NCQA). Examples of primary care physician measures that health plans use as part of physician reimbursement and reporting include cholesterol monitoring for patients with cardiovascular conditions; depression screening; and cervical cancer, breast cancer, and colon cancer screenings. There are also measures that reward providers for increased use of generic prescriptions as opposed to brand pharmaceuticals that are more expensive.

Quality Outcomes Measures

Ideally, health plans should want to measure, report, and pay based on quality outcomes alone (which physician improved his or her patients' health). However, there is no consensus as to how one would measure this practice. Each provider can claim that his or her patients are sicker and therefore cannot be compared to results with the rest of the population. Instead, focus measures for physicians are the process, efficiency, and technology measure. On the hospital side, there are some outcome measures that most health plans use to measure and reimburse hospitals based on their performance. The Agency for Healthcare Research & Quality (AHRQ) has a prominent role in establishing hospital measures that are based on administrative claims data that hospitals submit to the states in which they operate. AHRQ measures include rates of hospital-acquired pneumonia and obstetrical trauma during child delivery.

Technology Adoption

Health plans logically recognize that the use of technology such as electronic prescription, computer physician order entry, and electronic medical records should lead to safer care, better outcomes, and less medical errors. For this reason, health plans often offer bonus payments for technology adoption and will help to supplement payment for that technology based on the assumption that it will lower costs and improve quality outcomes.

Consumers

Principles held by consumers include:

- Stay healthy
- Affordable healthcare
- No hassles from insurer
- Choose doctors based on recommendations of family or friends
- Have trustworthy physicians
- Free from medical errors

Quality Information as Defined by Consumers

Consumers are increasingly using the Internet to look up symptoms, conditions, and treatment options; they often print this information and will use it as a basis for discussion with their physicians. They should also have information on how to compare and choose a physician or surgeon. They should be able to choose which physician is best for them based on quality outcomes. Consumers tend to care about the following criteria:

- **Information on access to care:** How close is the physician and hospital to where they live, and how long will they have to wait for an appointment?
- **Information on patient experience:** How satisfied are past patients, and therefore, what is the level of their satisfaction with this physician or hospital?
- **Information on value:** This is especially important for members who have to pay a significant portion of their care out of pocket. They need to decide if it is worth going to a particular physician and getting a particular treatment.
- **Information on similar experiences:** Patients want to know how the provider treated patients similar to them in terms of conditions and health status. Consumers also believe indicators of quality include the physician's training, age, gender, and nationality. As mentioned previously, there is very limited publicly available data on physicians, but more on the hospitals.

- **Information on lawsuits:** Patients want to know if a physician has any past or current malpractice suits.

Employers

Principles held by employers:

- Control costs
- Keep employees healthy and working (reduce absenteeism)
- Understand the medical needs by employees

Quality Information Defined by Employers

Large employers especially want to be able to track (within privacy restrictions) the conditions and diseases most prevalent among their workforce and devise how they should best be treated to keep these employees healthy and working, while keeping costs under control. Employers want to track absenteeism related to medical care to keep their employees as productive as possible.

In 1998, more than 170 large employers such as GE and Ford began to come together on the issue of hospital-patient safety. Members of Leapfrog Group, a coalition of employers, agreed to influence hospitals to institute programs and health plans to reward providers that meet patient safety standards. To further its mission, Leapfrog has developed an incentive program that it licenses to health plans to kick off their pay-for-performance programs. On the physician side, Bridges to Excellence, a coalition of employers, payers, and providers, licenses their physician incentive quality program to health plans such as United, Cigna, and Carefirst. Leapfrog Group and Bridges to Excellence have strong support in the employer community.

Government

Principles held by government:

- Control healthcare costs and spending
- Support electronic medical records
- Improve quality
- Pay for better quality (new)

Quality Information as Defined by Government

- Health outcomes. This includes volume of procedures, mortality, complications, and satisfaction
- Technology adoption, as it is assumed to lead to safer and more efficient care
- Cost efficiency

In 2004, the Center for Medicare Services (CMS) launched a voluntary reporting program for hospitals with a starter set of 10 quality measures related to three medical conditions: heart attack, heart failure, and pneumonia. The list of measures has since increased to 34 and includes measures for bypass surgery and hip and knee replacements. The program is still voluntary, but hospitals that do not report take a .4% reduction in CMS reimbursement. CMS is conducting a pilot financial bonus program that rewards the top 20% of participating pilot hospitals with a 1% to 2% bonus. Premier is one of the pilot sites and reports 10% improved patient care since beginning the program.

Some states have begun to require mandatory reporting to consumers on physicians and hospitals. Minnesota has passed a law requiring that physicians and hospitals are not reimbursed for "never events"—egregious medical errors like wrong side surgery or surgery on the wrong patient for the wrong reason. For now, each state is defining these "never events," and policies vary across states and payers.

FUTURE TRENDS

The healthcare marketplace is moving fast toward digitizing information. Software vendors provide enterprise versions of databases that can be used by physicians and nurses that are linked to patients' records. Claims data are also moving toward including more comprehensive clinical information and being accessible in a timely manner. Currently, there is up to a one- to two-year lag in the payment after the actual performance period occurrence. As more physicians and hospitals implement electronic medical records, there needs to be open discussion about what data need to be captured and what security interface mechanisms should be established so that relevant quality outcomes can be shared across the healthcare system without jeopardizing patient privacy and confidentiality. Electronic records improve the information collection time and information accessibility from its reusability. However, we also need to understand that poor quality information can spread even faster with electronic records.

Currently, numerous systems are being developed and deployed. Many systems are not well integrated to share data. Particularly, integrating financial data and medical data is an issue to resolve. Increasingly, the issues of privacy and confidentiality have become an important aspect of quality information that critically deserves a resolution.

CONCLUSION

Improving information quality is an idea-in-good-currency in the healthcare marketplace. Many hospitals and health plan providers are engaged in ways to provide quality information. The demand for quality information from the consumers pushed hospitals and health plan providers to spend more resources in restructuring their internal information quality managers and analysts to address the issue in a continuous and integrated way.

Quality of information in the healthcare marketplace is considered to be a critical determinant of the quality of medical service delivered. With the emphasis on evidence-based medicine, information is the asset for various analyses geared at finding out efficient and effective ways of using resources for quality care. With elevated discussion on quality information, the various concerns by patients, employers, physicians, hospitals, and health plan providers will have a better chance to be addressed. It is time to act on establishing information quality strategy (Lee et al., 2006) to provide quality healthcare information.

REFERENCES

Aas, I. (2001). A qualitative study of the organizational consequences of telemedicine. *Journal of Telemedicine and Telecare, 7*(1), 18–26.

Bashshur, R., Reardon, T., & Shannon, G. (2001). Telemedicine: A new healthcare delivery system. *Annual Review of Public Health, 21*(1), 613–637.

Berwick D. (1995). *Quality management in healthcare, 4*(1), 27–55.

Butterfield, K. (2004). An inductive model of collaboration from the stakeholder's perspective. *Business & Society, 42*(2), 162–195.

Davidson, B., Lee, Y., & Wang, R. (2004). Developing data production map: Meeting patient discharge submission requirement. *International Journal of Healthcare Technology and Management, 16*(2), 87–103.

Freeman, R., & Reed, D. (1983). Stockholders and stakeholders: A new perspective on corporate governance. *California Management Review, 25*(3), 88–106.

Gagnon, M., et al. (2005). Telehealth adoption in hospitals: An organizational perspective. *Journal of Health Organization and Management, 19*(1), 32–56.

Hu, P. Chau, P., Liu, O., & Tam, K. (1999). Examining the technology acceptance model using physician acceptance of telemedicine technology. *Journal of Management Information Systems, 16*(2), 91–112.

Jayasuriya, R. (1998). Determinants of microcomputer technology use: Implications for education and training of health staff. *International Journal of Medical Informatics, 50*, 187–194.

Katz-Haas, & Lee, Y. (2005). Understanding interdependencies between information and organizational processes. *Advances in Management Information Systems, 11*, 167–180.

Kohn, K., Corrigan, J., & Donaldson, M. (1999). *To err is human: Building a safer health system*. Washington, D.C.: National Academy Press.

Kohn, L., Corrigan, J., & Donaldson, M. (1999). The Institute of Medicine Report Committee on quality of healthcare in America. Institute of Medicine.

Landro, L. (2004). Booster shot: To get doctors to do better, health plans try cash bonuses—"Pay for performance" rewards preventative, follow-up care; physicians resist checkup—"No one cook book solution." *The Wall Street Journal*, September 17.

Leape L., & Berwick, D. (2005). Five years after to err is human. *JAMA, 293*(19).

Learning and Information Services, University of Hertfordshire. Retrieved January 10, 2006, from http://www.herts.ac.uk/lis/subjects/health/ebm.thm

Leatherman, S., & Berwick, D. (2003). *Health Affairs*, 03-2003.

Lee, Y. (2004). Crafting rules: Context-reflective data quality problem solving. *Journal of Management Information Systems, 20*(3), 93–119.

Lee, Y., & Pipino, L. (2004). Unpublished internal interview.

Lee, Y., Pipino, L., Funk, J., & Wang, R. (2006). *Journey to data quality*. Cambridge, MA: MIT Press.

Lowenstein, R. (2005). The quality cure? *New York Times*, March 13.

Med-Vantage, Inc. (2005). Provider-pay-for-performance incentive programs, first quarter.

Menon, M., Lee, B., & Eldenburg, L. (2000). Productivity of information systems in the healthcare industry. *Information Systems Research, 11*(1), 83–92.

Powell, J. (2005). Hospital view: Time to scrub. *Boston Herald*, December 23, 27.

Raitoharju, R., & Laine, M. (2006). *Exploring the differences in information technology acceptance between healthcare professionals*. Proceeding of the Americas Conference on Information Systems (AMCIS), Acapulco, Mexico.

Rosenthal, M. (2004). Paying for quality: Provider incentives for quality improvement. *Health Affairs, 23*(2), 212–232.

Skidmore, S. (2005). A growing practice—Programs that compensate doctors based on performance are becoming more common, *Signon*. Retrieved December 20, 2005, from www.SignOnSandiego.com

Strong, D., Lee, Y., & Wang, R. (1997). Data quality in context. *Communications of the ACM, 40*(5), 103–110.

Traynor, K. (2005). Study finds modest benefit in pay for performance. *American Society of Health-System Pharmacists*. Retrieved December 15, 2005, from http://www.ashp.org/news/ShowArticle.cfm

Wang, R., Lee, Y., Pipino, L., & Strong, D. (1998). Manage your information as a product. *Sloan Management Review, 39*(4), 95–105.

Wang, R., & Strong, D. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5–33.

Wennberg, J. (2006). The care of patients with severe chronic illness: An online report on the Medicare program. *Dartmouth Atlas Project Report*, May 16.

KEY TERMS

Quality Care: Care based on evidence-based guidelines that seeks to address overuse, misuse, and underuse services (Kohn, Corrigan & Donaldson, 1999).

Quality Information: Information that is fit for use by information consumers.

A Historical Overview of Health Disparities and the Potential of E-Health Solutions

Michael Christopher Gibbons
Johns Hopkins Urban Health Institute, USA

INTERNATIONAL ORIGINS

Over the past decade, a rapidly expanding body of scientific evidence has been put forth documenting differences in health status among U.S. racial and ethnic groups. Evidence has also mounted suggesting that these differences may be related to both medical and nonmedical determinants. Internationally, however, neither the evidence nor the realization of a link between nonmedical sociobehavioral factors and health outcomes is new. The earliest reported observation of a hypothesized association between socioenvironmental risk factors and health outcomes occurred in Italy over three centuries ago when Bernardino Ramazzini detailed an unusually high frequency of breast cancer in Catholic nuns (Wilson, Jones, Coussens, & Hanna, 2002). Not long thereafter, in 1775, British surgeon Sir Percival Pott reported a cluster of scrotal cancer cases among British chimney sweeps (Wilson et al., 2002).

By the mid-19th century, large-scale epidemiologic evidence began to corroborate these early observations. In 1840, Edwin Chadwick, British civil servant and statistician, demonstrated mortality differentials between the social classes living in Liverpool, England. Chadwick asserted that these differences were likely due to poverty and lifestyle factors common to the poorer working classes (Macintyre, 1997). German physician Rudolph Virchow went a step further when, in 1849, he asserted that because diseases of the populace are traceable to defects in society, the focus of medicine should shift from changing the individual to that of changing the society (Amick, Levine, Tarlov, & Walsh, 1995). Finally, in France, French physician Louis Villerme recommended improving school and working conditions as social interventions that would reduce class differences in mortality (Amick et al., 1995). Thus, in Europe, by the beginning of the 20th century, the existence of class variations in morbidity and mortality were clearly evident in the scientific literature (Macintyre, 1997).

Throughout the 20th century, the study of social class differences in health status continued across Europe, especially in Britain where epidemiologists began using decennial census data to evaluate national mortality trends. The insights gained from these analyses enabled them to construct an occupational social class grading system that correlated inversely with infant mortality. It also was the basis of the claim made by the Registrar General of Britain that at least 40% of British infant mortality was entirely preventable if the social conditions of poor infants could be elevated to that of upper-class infants (Macintyre, 1997).

Two British researchers, Titmuss and Logan, evaluated regional class-based mortality trends and documented that the disparity in infant mortality rates between upper- and lower-class infants continued to increase from 1910 to 1950 (Macintyre, 1997). This data, along with the Depression and World War II, encouraged the British government in 1942 to respond by instituting the welfare state and promoting several policy initiatives designed to address the "five giants of Want, Disease, Ignorance, Squalor and Idleness" (Acheson, 1998; Macintyre, 1997). Despite this government investment, however, problems attributable to social inequalities and inadequate access to health care persisted. In fact, by the mid-1970s, some 30 years later, the evidence seemed to indicate that the problems were still increasing and that the health of British citizens was slipping behind that of other industrialized nations (Acheson, 1998). Thus, in 1977, the British government formed the Research Working Group on Inequalities in Health and selected Sir Douglas Black as its chair. The committee's report, issued three years later in 1980, became known as the Black Report, and it represents the first attempt by a national government to systematically study, understand, and explain health inequalities (Acheson, 1998). In summary, the health improvement recommendations of the report emphasized the need to improve the physical and social environments in which the poor and lower classes lived (Acheson, 1998).

DOMESTIC RECOGNITION

Across the Atlantic in the United States, scientific evidence from several lines of inquiry examining outcomes and patterns of health care delivered to defined populations began to converge and suggest the importance of the socioenvironment in determining health outcomes. Researchers using small area analysis and geographic information systems analytic techniques demonstrated that a significant amount of nonrandom practice variability existed between clinical practices in different geographic locales, despite treating clinically similar patients (Barnes, O'Brien, Comstock, D'Arpa, & Donahue, 1985; McPherson, Wennberg, Hovind & Clifford, 1982). As public awareness grew, the U.S. government became involved. In 1984, the U.S. Department of Health and Human Services released a report on the health of the nation, titled *Health, United States, 1983* (NCHS, 1983). The report documented that while the overall health of the nation showed significant progress, major disparities existed in "the burden of death and illness experienced by blacks and other minority Americans as compared with the nation's population as a whole" (NCHS, 1983).

In response to the disparities identified in the report, the secretary of the Department of Health and Human Services established a task force on black and minority health—the first time that the U.S. government formed a group of experts to conduct a comprehensive study of minority health problems. In 1985, release of the "Report of the Secretary's Task Force on Black and Minority Health" significantly raised awareness of the disparate health of the country's minority groups compared to the white majority population (Mayberry, Mili, & Ofili, 2000).

Large epidemiologic studies like the Harvard Medical Practice Study emerged, documenting that a significant portion of practice variability could be classified as substandard care and that there was a correlation between substandard care and health care centers treating substantial numbers of poor and minority patients (Brennan, Leape, Laird, Localio & Hiatt, 1990; Brennan et al., 1991; Leape et al., 1991).

The emerging problems of differential outcomes and health status were not limited, however, to minorities and the poor. The Whitehall studies of a large cohort of British civil servants had convincingly demonstrated that a social class-based health gradient existed even among the well educated and employed (Marmot, Rose, Shipley & Hamilton, 1978). Additionally, it became increasingly recognized that certain community and societal level factors, including stress (Marmot, 1986; Sapolsky & Mott, 1987), early life experiences (Tager, Weiss, Munoz, Rosner, & Speizer, 1983), social capital (Coleman, 1988), and income inequality (Wilkinson, 1992a, 1992b) seemed to exert significant effects on health and disease outcomes independent of personal behavior (Amick et al., 1995; Brennan et al., 1990; Wilkinson, 1996). Soon, major philanthropic and advocacy organizations, including the Commonwealth Fund, the Kaiser Family Commission, the Kellogg Foundation, the Robert Wood Johnson Foundation, and the California Endowment began major initiatives designed to address issues related to disparities and health care quality (Compendium of Cultural Competence Initiatives in Healthcare, 2003a).

By the late 1990s, the scientific evidence seemed to indicate that issues of disparity, practice variation, substandard care, and socioenvironmental determinants of health may all be related to the quality of health care experienced by patients. Fiscella, Franks, Gold, and Clancy (2000) published a paper titled "Inequality in Quality," in which they called attention to issues of health care quality and health care disparities as related issues of health care organizational capacity. They further contended that national efforts to eliminate racial and ethnic disparities in health care and national health care quality improvement initiatives represented two inseparable components of providing high-quality health care for all citizens (Fiscella et al., 2000).

SYNTHESIZING THE SCIENTIFIC EVIDENCE ON HEALTH DISPARITIES

As the domestic evidence for population differences continued to accumulate, definitions of disparities were nonstandardized and racial categorizations became increasingly criticized as being imprecise and biologically meaningless (Anderson & Nickerson, 2005; Smedley & Smedley, 2005). While multiple definitions are still in current use, disparities are generally held to be population differences (Wilson et al., 2002) in environmental exposures (Macintyre, 1997) health care access, utilization, or quality, (Amick et al., 1995) health status (Acheson, 1998), or health outcomes (Carter-Pokras, & Baquet, 2002). As alluded to previously, within the U.S. health care system, these differences have most convincingly been demonstrated across racial and ethnic lines (whites vs. minorities); however, disparities based on other categorizations have also been described, including geography (urban vs. rural) (2005), gender (male vs. female) (McGrath & Puzan, 2004; Quinn & Overbaugh, 2005), socioeconomic status (poor vs. nonpoor) (Adler & Ostrove, 1999; Federico & Liu, 2003), and age (nonelderly vs. elderly) (Pyle & Stoller, 2003).

Health disparities are generally thought to be related to the health care system and other social factors. Several lines of investigation examining the socioenvironment and the clinical encounter give evidence of differences in the quality of care received by many racial and ethnic minorities. While these factors have been described as "causes" and are likely to be important in the genesis of disparities, scientifically validated evidence of definitive causal pathways and the underlying biologic mechanisms is largely lacking (Evans, Barer, & Marmor, 1994).

To help bring clarity to these issues, the Institute of Medicine released the first of several reports highlighting and summarizing the scientific evidence concerning issues of differential health status, culture, behavior, communication, substandard care/medical errors, and health care quality (2002; Smedley, Stith, & Nelson, 2003; 2001; IOM Committee on Quality of Healthcare in America, 2001; 2003b; Kohn, Corrigan, & Donaldson, 2000; Haynes & Smedley, 1999). The work of the IOM on disparity issues culminated with the 2003 release of a report titled "Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care" (Smedley et al., 2003). In this report, the IOM Committee on Understanding and Eliminating Racial and Ethnic Disparities in Health Care was charged with assessing the extent and potential sources of racial and ethnic disparities in health care that are not otherwise attributable to access to care, ability to pay, or insurance coverage. The committee was also to provide recommendations regarding potential interventions to eliminate health care disparities (Smedley et al., 2003). The committee found that within the United States, even among individuals with access to care, significant racial and ethnic disparities indeed existed and were related to historic and contemporary social and economic inequality, discrimination, and a fragmented U.S. system of health care (Smedley et al., 2003). While the release of this report has engendered significant public, media, and academic interest, likely ensuring that efforts to understand and eliminate disparities will continue at least into the foreseeable future, the magnitude and intransigence of the problem, the complexity of its causal pathways, and its resistance to intervention efforts are only beginning to be realized (2002).

DIGITAL DISPARITIES

Since the mid-1990s when the World Wide Web became a powerful part of America's communications and information culture, there has been great concern that the nation's racial minorities would be further disadvantaged because Internet access was not spreading as quickly in the African-American community as it was in the white community. Former Assistant Secretary of Commerce Larry Irving said the following in his introduction to "Falling Through the Net," the 1999 Department of Commerce Study on the digital divide (the divide between those with access to new information technologies and those without): "[The digital divide] is now one of America's leading economic and civil rights issues" (US Department of Commerce, 1999a). This report found that although overall the number of Americans connected to the nation's information infrastructure was soaring, a digital divide existed between whites and African-Americans in terms of their access to the Internet, and that in many cases, the divide was *widening* over time. A follow-up study revealed a persistent but substantially narrowed gap, with large increases in computer ownership and Internet use across most major demographic populations (Spooner & Rainie, 2005). The most recent survey, released in 2003, indicated a significant slowing in the growth of the number of Internet users since late 2001 (Lenhart, Horrigan, Allen, Boyce, Madden, & O'Grady, 2003). Overall, 42% of surveyed individuals did not use the Internet, and significant utilization differences remained according to race, education, income, and geography (urban vs. rural) (Lenhart et al., 2003). Generally, whites are more connected than African-Americans and Hispanics. Even at equivalent levels of income, African-Americans are less likely to be online than whites or Hispanics. In fact, over the period of this study (mid-2000 to mid-2002), the composition of the non-Internet user group did not change substantially (Lenhart et al., 2003). Interestingly, 56% of nonusers said they did not ever plan to go online and cited the cost of computers or Internet access, fear of fraud, credit card theft, or pornography as the major reasons for avoiding Internet use (Lenhart et al., 2003).

Recently, there has been a significant increase in the public availability of computers and Internet access at schools, public libraries, and workplaces (Office of Disease Prevention and Health Promotion, 2002). Thus, conclusions regarding the extent of a digital disparity based on data considering only home-based access may be limited. Despite this reality, Internet availability in the home is accepted as an important indicator of equitable access among population groups (Office of Disease Prevention and Health Promotion, 2002). In addition, access in public settings may be problematic because of computer monitoring in the workplace, privacy and confidentiality concerns, and the facilities' hours of operation. Because of the potentially sensitive nature of health-related uses of the Internet, access at home is thought to be essential (Office of Disease Prevention and Health Promotion, 2002).

Several studies have shown that access to the Internet correlates with income level and educational attainment (Office of Disease Prevention and Health Promotion, 2002; Spooner & Rainie, 2005; Lenhart et al., 2003; U.S. Department of Commerce, 1999a). As with racial and ethnic differences, Internet utilization is increasing in all income brackets. The largest increases are seen in the higher income categories. All things considered, household incomes above US$50,000 are positively associated with Internet utilization (Lenhart et al., 2003). Beyond socioeconomic issues, some researchers have speculated that African-Americans have had less access to the Internet because they participate to a greater degree in entertainment-oriented technologies like television rather than in information technologies. They argue that relatively high proportions of African-Americans use radio and television, but a relatively low proportion read newspapers (Spooner & Rainie, 2005). As already suggested, the primary reasons why some groups have less access to information technology and resources are related to geography, literacy, disability, local infrastructure requirements, and cultural differences (Eng, Maxfield, Patrick, Deering, Ratzan, & Gustafson, 1998), some of which are not easily overcome simply by increasing personal computer ownership. Even if equity in personal computer and Internet access were achieved, emerging evidence suggests that online habits may vary by race and ethnicity. For example, online African-Americans are more likely than online whites to have (1) searched for information about major life issues such as researching new jobs and finding places to live, (2) used entertainment online, (3) used the Internet to obtain health information, and (4) searched for religious or spiritual information [40]. On the other hand, African-Americans with access to the Internet do not go online as often on a typical day as whites do, and they do not participate on a daily basis in most Web activities at the same level as online whites (Spooner & Rainie, 2005).

As information technology plays an ever-increasing role in Americans' economic and social lives, the potential health implications of these findings need to be more clearly evaluated because the prospect that some people will be left behind in the information age may have serious repercussions (U.S. Department of Commerce, 1999b). Persistent digital disparities in access or utilization could leave some groups less able to take advantage of cutting edge innovations in population health technologies that enhance disease surveillance, environmental monitoring, food safety, emergency planning, disaster management, and geographic information systems-based tracking of environmental hazards (Eng, 2004).

THE ROLE OF INFORMATION TECHNOLOGY IN OVERCOMING HEALTH DISPARITIES

One major domain of eHealth focuses on improving health communication through the use of technology. This notion of enhancing communication and understanding is a fundamental component of addressing health disparities. Among other things, the recommendations of the Institute of Medicine report call for initiatives designed to enhance patient-provider communication, trust, and cultural appropriateness of delivered care (Smedley et al., 2003). Similar goals are the basis for the Healthy People 2010 objective to increase the number of individuals with Internet access in the home. Providers, health care organizations, and public health agencies are increasingly using the Internet as a main source of information dissemination and communication (Office of Disease Prevention and Health Promotion, 2002). This need for innovative improvements in communication should represent a significant opportunity for eHealth technologies, researchers, and interventionists, with many important implications for overcoming disparities in health and health care. Given that eHealth is currently understood as attempting to facilitate the utilization of information technologies, the Internet, and communication technology in order to facilitate behavior change, improve health care, and enhance health outcomes (Eng, 2002), eHealth researchers may become the catalysts needed to spur the development of transdisciplinary interventions to effectively address disparities in health and health care.

Recent advances in the computer sciences and information technology fields have spawned several methodological advances in the biological and molecular sciences (e.g., DNA chip technology and microarray analysis), enabled quantum leaps in molecular and submolecular medicine, and catalyzed the emergence of whole new fields of study such as proteomics, phenomics, nutrigenomics, and pharmacogenetics. Perhaps in like manner with the emergence of eHealth, the behavioral and population sciences may be on the verge of a similar information technology-based scientific revolution. New eHealth solutions may soon permit real-time integrative utilization of vast amounts of behavioral-, biological-, and community-level information in ways not previously possible. Behavioral algorithms and decision support tools for scientists could facilitate the analysis and interpretation of population level data to enable the development of "community (population) arrays" or communitywide risk profiles, which in turn could form the foundation of a new "populomics." This population-level risk characterization could potentially go beyond the limitations of typical geographic analyses and yield insights distinctly different from risk stratification based on current methodologies. Generically, these emerging technologies have been termed population health technologies and are believed to offer significant promise (Eng, 2004).

These assertions are not based on mere speculation. Encouraging early evidence suggests that multimedia health communication and behavior change efforts that include the use of computers and other eHealth technologies can improve health outcomes (Neuhauser & Kreps, 2003). Among other factors, the evidence suggests that applications that are tailored to the individual, participatory, personally relevant, and contextually situated will be more likely to promote behavior change (Neuhauser & Kreps, 2003). On the other hand, the Internet has been implicated in the causation or persistence of disparities because of the relative lack of access of some groups and because of its current inability to deliver content that is dynamically tailored to meet the cultural, language, or literacy needs of the individual user (Cashen, Dykes, & Gerber, 2004). This may be particularly true of eHealth applications that are Internet-enabled, requiring access to the Internet to provide the interventional content. It is conceivably possible, however, to conceptually divide eHealth applications into at least two genres: those that rely on the Internet to deliver the interventional content directly to patients (Internet-enabled), and those that only employ the Internet to facilitate transfer and utilization of data for or about content that is delivered to patients by an alternate approach. The content or interventions themselves can actually function without the Internet, but when used in the context of the Internet, they are potentially much more efficacious and far-reaching. These types of technologies could be termed "Internet-enhanced" eHealth solutions. Here, the Internet would facilitate the transfer of data and information, but the tailored content could be delivered by trusted people from the user's own culture or community. The actual intervention could also be administered to patients by print or multimedia applications. Thus, in terms of overcoming health disparities, issues of guaranteeing Internet access for every individual may prove to be less important than attempting to address health disparities via interventions and methodologies that lack cultural relevance. Indeed, those interventions and strategies that integrate behavioral interventions with emerging information technologies will likely be the interventions capable of cost-effectively enabling mass customization, interactivity, and convenience. Ultimately, though, the health disparities challenge for eHealth researchers remains to harness the technical capabilities of emerging information technologies in ways that support the social and cultural realities in which people work and live (Neuhauser & Kreps, 2003), while enhancing our ability to address the health needs of every patient (Safran, 2003).

ACKNOWLEDGMENT

The author wishes to acknowledge Earl Fox for his support of the author's work in this area.

REFERENCES

Acheson, D., Barker, D., Chambers, J., Graham, H., Marmot, M., & Whitehead, M. (1998). *Independent inquiry into inequalities in health report*. The Stationery Office. UK Department of Health.

Adler, N.E., & Ostrove, J.M. (1999). Socioeconomic status and health: What we know and what we don't. *Ann NY Acad Sci, 896,* 3–15.

Amick, B.C., Levine, S., Tarlov, A.R., & Walsh, D.C. (1995). *Society and health*. New York: Oxford University Press.

Anderson, N.B., & Nickerson, K.J. (2005). Genes, race, and psychology in the genome era: An introduction. *Am Psychol, 60,* 5–8.

Barnes, B.A., O'Brien, E., Comstock, C., D'Arpa, D.G., & Donahue, C.L. (1985). Report on variation in rates of utilization of surgical services in the Commonwealth of Massachusetts. *JAMA, 254,* 371–375.

Brennan, T.A., et al. (1991). Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I. *N Engl J Med, 324,* 370–376.

Brennan, T.A., Leape, L.L., Laird, N.M., Localio, A.R., & Hiatt, H.H. (1990). Incidence of adverse events and negligent care in hospitalized patients. *Trans Assoc Am Physicians, 103,* 137–144.

Carter-Pokras, O., & Baquet, C. (2002). What is a "health disparity"? *Public Health Rep, 117,* 426–434.

Cashen, M.S., Dykes, P., & Gerber, B. (2004). eHealth technology and Internet resources: Barriers for vulnerable populations. *J Cardiovasc Nurs, 19,* 209–214.

Coleman, J.S. (1988). Social capital in the creation of human capital. *Am J Sociol, 94,* S95–120.

Compendium of cultural competence initiatives in healthcare. (2003a). Washington, DC: The Henry J. Kaiser Family Foundation.

Eng, T. R. (2002). eHealth research and evaluation: Challenges and opportunities. *J Health Commun, 7,* 267–272.

Eng, T.R. (2004). Population health technologies: Emerging innovations for the health of the public. *Am J Prev Med, 26,* 237–242.

Eng, T.R., Maxfield, A., Patrick, K., Deering, M.J., Ratzan, S.C., & Gustafson, D.H. (1998). Access to health information and support: A public highway or a private road? *JAMA, 280,* 1371–1375.

Envisioning the national health care quality report. (2001). Washington, DC: National Academy Press.

Evans, R.G., Barer, M.L., & Marmor, T.R. (1994). *Why are some people healthy and others not? The determinants of population health.* New York: Aldine De Gruyter.

Federico, M.J., & Liu, A.H. (2003). Overcoming childhood asthma disparities of the inner-city poor. *Pediatr Clin North Am, 50,* 655–675, vii.

Fiscella, K., Franks, P., Gold, M.R., & Clancy, C.M. (2000). Inequality in quality: Addressing socioeconomic, racial, and ethnic disparities in health care. *JAMA, 283,* 2579–2584.

Handbook of urban health, populations, methods and practice. (2005). New York: Springer.

Haynes, M., & Smedley, B.D. (1999). *The unequal burden of cancer*. Washington, DC: National Academy Press.

IOM Committee on Quality of Healthcare in America. (2001). *Crossing the quality chasm: A new health system for the 21st century*. Washington, DC: National Academy Press.

Kohn, L., Corrigan, J., & Donaldson, M. (2000). *Too err is human: Building a safer health system*. Washington, DC: National Academy Press.

Leape, L.L., et al. (1991). The nature of adverse events in hospitalized patients. Results of the Harvard Medical Practice Study II. *N Engl J Med, 324,* 377–384.

Lenhart, A., Horrigan, J., Allen, K., Boyce, A., Madden, M., & O'Grady, E. (2003). *The ever-shifting Internet population: A new look at Internet access and the digital divide.* Washington, DC: Pew Charitable Trusts.

Macintyre, S. (1997). The Black Report and beyond: What are the issues? *Soc Sci Med, 44,* 723–745.

Marmot, M.G. (1986). Does stress cause heart attacks? *Postgrad Med J, 62,* 683–686.

Marmot, M.G., Rose, G., Shipley, M., & Hamilton, P.J. (1978). Employment grade and coronary heart disease in British civil servants. *J Epidemiol Community Health, 32,* 244–249.

Mayberry, R.M., Mili, F., & Ofili, E. (2000). Racial and ethnic differences in access to medical care. *Med Care Res Rev, 57*(Suppl 1), 108–145.

McGrath, B.B., & Puzan, E. (2004). Gender disparities in health: Attending to the particulars. *Nurs Clin North Am, 39,* 37–51.

McPherson, K., Wennberg, J.E., Hovind, O.B., & Clifford, P. (1982). Small-area variations in the use of common surgical procedures: An international comparison of New England, England, and Norway. *N Engl J Med, 307,* 1310–1314.

NCHS. (1983). *NCHS: Health, United States, 1983; and prevention profile* (Rep. No. (PHS) 84-1232). Washington, DC: US Government Printing Office.

Neuhauser, L., & Kreps, G.L. (2003). Rethinking communication in the e-health era. *Journal of Health Psychology, 8,* 7–22.

Office of Disease Prevention and Health Promotion. (2002). *Healthy people 2010, 2nd Ed.* Washington, DC: Government Printing Office.

Priority areas for national action: Transforming health care quality (2003b). Washington, DC: National Academy Press.

Pyle, M.A., & Stoller, E.P. (2003). Oral health disparities among the elderly: Interdisciplinary challenges for the future. *J Dent Educ, 67,* 1327–1336.

Quinn, T.C., & Overbaugh, J. (2005). HIV/AIDS in women: An expanding epidemic. *Science, 308,* 1582–1583.

Safran, C. (2003). The collaborative edge: Patient empowerment for vulnerable populations. *Int J Med Inform, 69,* 185–190.

Sapolsky, R.M., & Mott, G.E. (1987). Social subordinance in wild baboons is associated with suppressed high density lipoprotein-cholesterol concentrations: The possible role of chronic social stress. *Endocrinology, 121,* 1605–1610.

Smedley, A., & Smedley, B.D. (2005). Race as biology is fiction, racism as a social problem is real: Anthropological and historical perspectives on the social construction of race. *Am Psychol, 60,* 16–26.

Smedley, B.D., Stith, A.Y., & Nelson, A.R. (2003). *Unequal treatment: Confronting racial and ethnic disparities in healthcare*. Washington, DC: National Academies Press.

Speaking of health: Assessing health communication strategies for diverse populations. (2002). Washington, DC: National Academy Press.

Spooner, T., & Rainie, L. (2005). *African-Americans and the Internet*. Washington, DC.

Tager, I.B., Weiss, S.T., Munoz, A., Rosner, B., & Speizer, F.E. (1983). Longitudinal study of the effects of maternal smoking on pulmonary function in children. *N Engl J Med, 309,* 699–703.

U.S. Department of Commerce, N.T. a. I.A. (1999a). *Falling through the net: Defining the digital divide*. Washington, DC: GPO.

U.S. Department of Commerce, N.T. a. I.A. (1999b). The digital divide summit. http://www.ntia.doc.gov/ntiahome/digitaldivide/summit/

Wilkinson, R.G. (1992a). Income distribution and life expectancy. *BMJ, 304,* 165–168.

Wilkinson, R.G. (1992b). National mortality rates: The impact of inequality? *Am J Public Health, 82,* 1082–1084.

Wilkinson, R.G. (1996). *Unhealthy societies: The afflictions of inequalities*. New York: Routledge.

Wilson, S., Jones, L., Coussens, C., & Hanna, K. (2002). *Cancer and the environment: Gene-environment interaction*. Washington, DC: National Academy Press.

KEY TERMS

Access to Information: Ability to get the right information at the right place at the right time.

Digital Disparities: Differences in groups of populations due to impact of technology.

E-Health: Health care information or services provided using Web-based technologies.

Internet: Web-based technologies.

Socioeconomic Factors: Social and economic factors that impact the ability to get needed health care solutions.

How Can Human Technology Improve the Scheduling of Unplanned Surgical Cases?

Janna Anneke Fitzgerald
University of Western Sydney, Australia

Martin Lum
Department of Human Services, Australia

Ann Dadich
University of Western Sydney, Australia

INTRODUCTION

Human technology in health care includes managerial knowledge required to marshal a health care workforce, operate hospitals and equipment, obtain and administer funds, and, increasingly, identify and establish markets. In this article, the authors focus on human technology and improvement of decision-making processes in the context of operating theatre scheduling of unplanned surgical cases.

Unplanned surgery refers to unscheduled and unexpected surgical procedures in distinction to planned, *elective* surgery. The management of unplanned surgery is a strategic function in hospitals with potential clinical, administrative, economical, social, and political implications. Making health care management decisions is complex due to the multidisciplinary and the multifocussed nature of decision-making processes. The complexity of multidisciplinary and multifocussed decision-making is further exacerbated by perceived professional identity differences.

This article presents findings from interviews with doctors and nurses about the scheduling of unplanned surgical cases. The interviews focused on current decision-making determinants, the acceptability of using a model to guide decision-making, and enablers and barriers to implementing the model. The key finding was the limited practicality of a model to guide the scheduling of unplanned surgery. While it could guide decisions around clinical determinants, logistical determinants, and ideal timeframes, it would have difficulty reshaping inter- and intra-professional dynamics.

LITERATURE REVIEW

The scheduling of unplanned surgery typically involves negotiating (and renegotiating) established surgery lists, whereby patients requiring nonelective surgery are attended to before those requiring elective surgery (Gabel, Kulli, Lee, Spratt, & Ward, 1999). However, delayed patient access to unplanned surgery can inflate economic costs (Jestin, Nilsson, Heurgren, Påhlman, Glimelius, & Gunnarsson, 2005; Pollicino, Haywood, & Hall, 2002). Poor patient health necessitates more health care services, including prolonged hospital stays, ongoing access to health care professionals, and medication. The tension between *health care cost containment* and the *high cost of surgical operations* is a powerful incentive for health care organisations to improve the management of the surgical suite (Gabel et al., 1999). Thus, an investigation of current scheduling practices of unplanned surgical cases is well-justified.

Empirical research on the scheduling of unplanned surgery is limited. This is somewhat reflected in the *ad hoc* practices found in some operating rooms when organising surgical queues (Fitzgerald, Lum, & Kippist, 2004). In Canada for instance, it was found that the waiting times for elective surgery were not only determined by the number of patients on the waiting list, or by how urgently they required treatment, but also by the *management* of the waiting list (Western Canada Waiting List Project, 2001). To ensure consistent practice, criteria were developed to guide decision-making processes. However, clinicians who managed the waiting lists were somewhat reluctant to change their management practices, preferring to adhere to less-standardised, conventional methods (Martin & Singer, 2003).

In the United Kingdom, surgical lists are typically compiled in an unplanned manner, and the negotiations and modifications that follow are also extemporised (Hadley & Forster, 1993). Even when theatre lists *are* established, they are seldom observed, often because of the need to accommodate patients who require unplanned surgery (Ferrera, Colucciello, Marx, Verdile, & Gibbs, 2001). This gives rise to extended surgery delays. Given such inconsistencies, the National Health Strategy (NHS) Executive circulated a national directive to guide good practice (Churchill, 1994). The directive emphasised the significance of operating room services, the prominence of patient care, the effective management and supervision of operating room use, staff morale, and communication, efficient transportation for patients, and dependable activity and cost information.

A standardised approach to manage operating room lists is also lacking in New South Wales (NSW), Australia (NSW Health, 2002a). This is not to suggest that NSW public hospitals lack direction for scheduling unplanned cases. Many have adopted priority codes to guide these decisions (NSW Health, 2002b). These help health care professionals to classify patients from highest to lowest priority based clinical need and/or timeframes.

However, research in NSW suggests that, in addition to clinical determinants, other factors influence the scheduling of unplanned surgery (Fitzgerald et al., 2004; King, Kerridge, & Cansdell, 2004); notably, logistics and inter- and intra-professional dynamics between staff involved in the process. Current policies for scheduling surgery, particularly unplanned surgery, fail to acknowledge these, providing little direction for accepted practice.

A review of literature on surgical decision-making practices suggests that the normative model of decision-making is *decision theory* (Gordon, Paul, Lyles, & Fountain, 1998; Magerlein & Martin, 1978; Parmigiani, 2002; Wright, Kooperberg, Bonar, & Bashein, 1996). Its ability to comprehensively consider information from diverse sources, especially in situations of great uncertainty, makes it particularly valuable.

Decision theory in operating theatres manifests through several models, including *queuing theory* (NSW Health, 1998) and the *Poisson distribution model* (Kim & Horowitz, 2002). Queuing theory commonly operates on a *first-come-first-served* basis, whereby priority is determined by chronology. However, in the case of unplanned surgery, where patient health outcomes are at-risk, this is illogical; and in the absence of explicit queuing rules, the theory causes confusion (Fitzgerald et al., 2004). In contrast, the Poisson distribution model operates with greater autonomy and is able to represent occurrences of a particular event—like unplanned surgery—over time or space. Using the Poisson distribution model for unplanned surgery would allow for the random arrival of patients, for assuming independence from other patient arrivals, and it supposes independence from the state of the hospital system (Clemen & Reilly, 2001). However, the model does not account for the multifactorial nature of clinical determinants and their relationship with time. Nor does it guide decisions in an environment where logistics and professional dynamics influence efficiency. Therefore, research into decision-making practices is needed to explore the tensions affecting the applicability of different queuing models.

Clinical Determinants and Time

Most operating rooms classify unplanned surgery according their relative urgency (Gabel et al., 1999). Although formalised classification systems are often lacking (NSW Health, 2002a), hospitals typically use a triage system to consider clinical determinants when scheduling both planned and unplanned surgery. Clinical decision-making practices therefore involve elements of rational choice theory (Johnson, 2000). It is deductive, based on evidenced-based practices that maximise patient health, and is therefore judicious.

However, there is no consistent system to allocate priority *between* unplanned cases (King et al., 2004). Furthermore, there is little quality auditing of the way clinical priorities are assigned. It is therefore possible that other factors determine the scheduling of unplanned surgery; critics of decision theory allude to this (Kahneman & Tversky, 1979; Redelmeier, Rozin, & Kahneman, 1993).

The unpredicted and often urgent nature of unplanned surgery suggests that *time* is an important factor when scheduling such surgery. Indeed some triage guidelines are entirely time-based, whereby certain clinical conditions must be done within a certain timeframe to prevent further loss of quality of life; there are thus serious time pressures within the operating room (Gabel et al., 1999) that may have direct consequence for clinical decision-making.

Logistical Determinants

Logistical determinants can further complicate the scheduling of unplanned surgery. A report concerning two public hospitals in Sydney, Australia, recognises the difficulties caused by limited resources and the ineffective management of operating theatres (King et al., 2004). For instance, it explains that while middle management personnel perform the executive management of the operating theatres, they are often absent or involved with other duties, leaving direct decision-making responsibilities to other, variably qualified staff. The report also mentions the physical fragmentation of operating theatre space; the management of large emergency and semi-urgent caseloads; and the inadequacy of systems to ensure that trauma patients are scheduled and prepared efficiently.

There are many more logistical determinants, including the availability of expert and support staff, specialist instruments, and theatre time and space. Therefore, despite the presenting *clinical* needs of the surgical patient, other factors can delay timely surgical intervention (Lebowitz, 2003). These in turn, have considerable economic implications (Bogner, 1997).

Professional Identity

Ineffective management is also partly consequent to surgeon desire to retain professional discretion (Buchanan & Wilson, 1996). Sociological literature has long recognised the domination of the medical profession (Freidson, 1970; Illich, 1977; Olesen, 2001). It is a profession characterised by *authority*, *autonomy*, and *sovereignty* (Bogner, 1997).

However, these attributes are not created and maintained by the medical profession in isolation; they are also influenced by the perceptions others have of the medical profession (Fitzgerald, 2002). These perceptions help to create strongly-defined professional boundaries that can divide different professional paradigms. Some of these boundary constructs include power, status, scientific educational background, clarity of task, and reward. The professional boundary between doctors and nurses is crystallised on each of these constructs.

Additionally, communication difficulties within and between professions are exacerbated by the lack of clear power structure between surgeons, anaesthetists, and nurse-managers within the operating theatre. This is particularly the case between surgeons and anaesthetists (Bogner, 1997; King et al., 2004).

Most operating room logistics are managed by the operating theatre manager, who is typically a nurse. This person is frequently the decision-maker who has a managerial perspective, incorporating the availability of beds and human resources with economic restraints, such as the staff-overtime budget. Together, these circumstances can create a climate of tension (Gabel et al., 1999). For those managers who use finesse and compromise to pursue a goal, confrontational behaviour from colleagues who demand perfectionism maybe difficult to manage. They may find themselves struggling to communicate appropriately, which, in turn, stifles the facilitation of effective management within the operating theatre.

Summary

A number of factors affect the scheduling of unplanned surgery. These include clinical determinants and time, logistical determinants and the dynamics between professionals. Because the complex nature of scheduling unplanned surgery can increase workplace frustration (Creswell, 2003; Stake, 1995), it is important to identify a framework that will facilitate effective decision-making practices in this area (Myers, 1997; Wilson, 2005). However, the implementation of this framework may pose difficulties.

To examine this possibility, this article presents findings from interviews with doctors and nurses, all of whom were presented with a decision-making model for scheduling unplanned surgical cases. The model was informed by previous research, which identified the clinical determinants, logistical determinants and ideal timeframes for scheduling unplanned surgical cases, according to doctors and nurses (Fitzgerald, Lum, & Dadich, 2006). Because the value of the model was largely unknown, it was crucial that its practicability be examined by those at the coalface.

METHOD

To recruit interviewees, the research team attended a statewide workshop at which 71 NSW operating rooms were represented. Operating room managers were invited to nominate their hospital as a possible research site. From the nominations, the researchers selected four metropolitan and four rural hospitals, each from different Area Health Services to limit geographical bias. Approval to conduct the research was gained from the university and Area Health Service ethics committees.

Hospital site coordinators were appointed for each participating hospital to assist with data collection. Hospital Site Coordinators were asked to recruit decision-makers in operating rooms, including surgeons, anaesthetists, nurse-managers, floor-managers, and administrators. Coordinators were then asked to invite interviewees to voluntarily contribute to the study.

A semistructured interview schedule was devised (and piloted) to explore the following themes:

- Current practices in the scheduling of unplanned surgery;
- The influence of clinical and time determinants;
- The influence of logistical determinants;
- The role of inter- and intra-professional dynamics when scheduling unplanned surgery;
- Methods to improve decision-making around the scheduling of unplanned surgery; and
- Thoughts about the proposed model (Fitzgerald et al., 2006).

Each interview was audio-taped and transcribed verbatim. QSR N-Vivo® software was used to aid detailed analysis of the data, facilitating the interpretation process (Creswell, 1998). This facilitated the emergence of core themes. To ensure consistency within each theme, codebooks were developed that included detailed descriptors of each theme, inclusion and exclusion criteria as well as exemplars from the research material.

RESULTS

Interviewees included surgeons (n=17), anaesthetists (n=9), and nurse- and nonclinical managers (n=12). There was much diversity in managerial qualifications—two (7%) doctors had undertaken internal management courses, while nine managers (83%) held qualifications ranging from certification to master degrees in management.

Corresponding with the literature review and the preceding discussion of clinical determinants and time, logistical determinants, and professional identity, the following section presents research findings that verify the influence of these factors on the scheduling of unplanned surgery. However, unlike existing literature, the findings suggest that the speed and efficiency of decision-making processes are mostly influenced by inter- and intra-professional dynamics.

Clinical Determinants and Time

Presented with the proposed model, the interviewees were asked to consider a taxonomy in which three clinical states were labelled *urgency 1*, *urgency 2,* and *urgency 3*. Collectively, they agreed that these categories represented conditions of *extreme urgency*, *intermediate urgency*, and *no urgency*. Within each category, there was recognition that specific physiological states were associated with different timeframes for surgical treatment:

The things that matter most are the urgency one and no one need a protocol to work out what's an urgency one… I have never seen an argument about that, everyone's in total agreement about its uniformity (surgeon).

It was evident that for categories other than *extreme urgency*, intuition was used to order the cases and times of surgery:

we got to feel it…. how things are likely to flow and give them our best guess (surgeon).

Although the three categories were statistically distinct, variations in acceptable responsiveness produces overlap and some ambiguity between the categories. Interviewees agreed that the overlap and ambiguity contributed to a "grey-zone":

when you got two cases that are very close, they probably would all come within more than one of these type of the categories (clinical care coordinator).

Although there was clear distinction between *extreme urgency*, *intermediate urgency*, and *no urgency*, there was ambiguity around the boundaries that separated each entity. Thus, five categories of clinical urgency might have the scope to accommodate individual case variation and overlap. A five-category model may help clinicians accept the need for dynamic decision-making. It also makes more explicit the clinical reasoning process for more transparent inputs, and provides common ground for reaching inter- and intra-professional agreement.

Logistical Determinants

Several logistical determinants influence the scheduling of unplanned surgery within NSW public hospitals, including the availability of space, theatre time, and staff, particularly surgeons. Because these are highlighted in relevant literature (Bogner, 1997; King et al., 2004; Lebowitz, 2003), focus will be given to findings seldom noted in the literature.

Inter- and Intra-Professional Dynamics

Most interviewees suggested that inter- and intra-professional dynamics influence the scheduling of unplanned surgery. These suggestions were often couched in observations about communication between personnel, which was sometimes cantankerous. This in turn, created clear distinctions between professions that are typically involved in the decision-making process—surgeons, anaesthetists and nurses. Furthermore, the comments suggest a strong hierarchy that sees surgeons having most decision-making power, followed by anaesthetists and nurses:

Some [surgeons] get very impatient and make a lot of noise… and we have used the anaesthetic department sometimes to adjudicate (nurse).

the consultant on-call has a power to veto [the decision to continue with planned surgery] and if you aren't happy with that decision, you call the consultant anaesthetist and see whether they'll… open an extra operating theatre… I guess… it comes down to… how hard the surgeon is prepared to push for the continuation of planned surgery, verses how hard the anaesthetic people are prepared to push for an extra theatre (anaesthetic registrar).

it's basically a situation where the nursing staff are really not making any decisions about patients; it's purely related to the medical team (surgeon).

Explicating these professional divisions, some interviewees noted the personal agendas that personnel bring to the decision-making process. Medical terms that lack an irrefutable definition are used to further individual interests:

surgeons tell fibs about what might be an emergency to suit themselves. That makes decisions difficult because they don't care if they categorise something a little bit differently, because they might have something to do elsewhere. They actually come out and admit that, but then they stick to their guns, that, "Yes this is an emergency" (nurse).

The hierarchical divisions between the three professions suggest concord *within* the professions. However, some interviewees recognised times of limited consensus among members of the *same* profession when attempting to schedule unplanned surgery.

I've seen conflict occur… There might be 100 surgeons of high-standing and ask them to assess the patient and you will get 100 different opinions about the diagnosis, let alone the urgency of treatment! (surgeon)

Occasionally, intra-professional disharmony was due to the different roles assumed by individual professionals. Some surgeons, anaesthetists, and nurses have accepted managerial roles within the hospital. With added responsibilities, they often bring an organisational perspective to the decision-making process. In the scheduling of unplanned surgery, they adopt a system-approach, ensuring the efficient use of existing resources:

We get caught up by nursing intensive care patients in recovery because there was no thought given to the aftercare the patient might require (manager).

Thus, a multifaceted decision-making process becomes more complex by managerial responsibilities that can (or should) constrain decisions.

The dual-identity of clinician-*cum*-managers can also generate intra-professional dissension. Non-managerial personnel are seen to "lack insight"; they are believed to have a limited understanding of the way decision-making processes affect the hospital as an organisation:

It is all very well to keep on providing services at all hours, at all costs. But we can't continue to do this in that way (manager).

Complex inter- and intra-professional dynamics suggest that models to guide decision-making practices can only be successful when there is opportunity to enhance these interactions.

DISCUSSION

This article demonstrates that while clinical state, time, and logistics influence the scheduling of unplanned surgical cases, equally important are inter- and intra-professional dynamics. Divergent views about the role of different professions depict an individualistic view of a context that requires collaborative decision-making. Surgeons in particular portray clear jurisdictions for clinical decision-making when scheduling unplanned surgery.

Nevertheless, there was some consensus between the different professions. Those interviewed recognised value in a model that guided the scheduling of unplanned surgery. The example presented to them resonated well, reflecting internal processes that are used intuitively (Fitzgerald et al., 2006). While some doctors and nurses warned against prescriptive and mechanistic protocols, surgeons, anaesthetists, and nurse-managers generally accepted a tool that provides common dialogue and facilitates effective decision-making.

The proposed model is thus a framework to bridge gaps in professional positioning. It allows dialogue to be redirected from individualistic drivers to greater collaboration.

Reflecting existing literature, the interviewees highlighted key determinants that influenced the scheduling of unplanned surgery; namely, clinical determinants and time (Gabel et al., 1999; Johnson, 2000), logistical determinants (Bogner, 1997; King et al., 2004; Lebowitz, 2003), and inter- and intra-professional dynamics (Buchanan & Wilson, 1996; Fitzgerald, 2002).

With regard to clinical determinants, the interviewees made clear distinctions between cases that are of *extreme urgency*, *intermediate urgency*, and *no urgency*. However, there was little agreement about where the boundaries lie that separate each category. Making decisions around the parameters of these three urgency classification may be assisted by a risk assessment model that incorporates reflection about *likelihood* and *consequence* of adverse events when delaying unplanned surgery.

Logistical determinants were also regarded as significant. The interviewees highlighted the availability of space, theatre time, and staff, particularly surgeons.

The intensity and volume of discussion on workplace interactions suggests that inter- and intra-professional dynamics impact more on the scheduling of unplanned surgery than current literature suggests. Dynamics between *and* within professions involved in this process can hinder the effective prioritisation of cases. Factors that thwart communication include the hierarchy that divides surgeons, anaesthetists, and nurses, personal agendas, difference of medical opinion *within* a profession, and the hybridity of roles assumed by clinician-*cum*-managers.

However, a number of methodological limitations must be considered. The cross-sectional nature of this study indicates that the interviewees provide a snapshot of the decision-making practices around the scheduling of unplanned cases within NSW public hospitals. Furthermore, qualitative research is limited by time, context, and the nature of individual perspectives (both the interviewee's and the researcher's interpretation). The use of convenience sampling to recruit interviewees may have also biased the findings.

FUTURE RESEARCH

Despite these limitations, the findings are consistent with those found in the existing body of relevant literature. However, most of this literature focuses on concrete triage determinants and little is published on inter- and intra-professional dynamics. Therefore, these findings pave the way for future research, particularly on the changing nature of professional jurisdictions in a climate where inter- and intra-professional decision-making is expected, as well as the organisational effects of hybrid roles.

CONCLUSION

In conclusion, this article confirms the dynamic and complex nature of decision-making around the prioritisation of unplanned surgery. It also suggests variability in the way unplanned cases are scheduled. In light of this, efforts to develop standardised practices through policy development may be somewhat futile, for they fail to appreciate the contextual differences of each hospital setting. Nonetheless, a decision-making tool, such as was presented to the interviewees, can improve decision making practices by acting as a catalyst for dialogue between and within professions when scheduling unplanned surgery.

REFERENCES

Bharadwaj, A., Sen, A., & Vinze, A. (1999). Scheduling cardiac procedures: A knowledge-based approach. *IEEE Transactions on Engineering Management, 46*(3), 322-334.

Bogner, M. S. (1997). Naturalistic decision making in health care. In C. E. Zsambok & G. A. Klein (Eds.), *Naturalistic decision making* (pp. 61-69). Mahwah, NJ: Lawrence Erlbaum Associates.

Buchanan, D., & Wilson, B. (1996). *Occasional paper 34: Re-engineering operating theatres: The perspective assessed*. Leicester: Leicester Business School.

Churchill, K. (1994). *Good practice in operating theatre management*. Leeds: NHS Executive.

Clemen, R. T., & Reilly, T. (2001). *Making hard decisions with decision tools*. Pacific Grove, CA: Duxbury.

Creswell, J. W. (1998). *Qualitative inquiry and research design: Choosing among five traditions*. Thousand Oaks, CA: Sage Publications.

Creswell, J. W. (2003). *Research design: Qualitative, quantitative, and mixed methods approaches* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Ferrera, P. C., Colucciello, S. A., Marx, J. A., Verdile, V. P., & Gibbs, M. A. (2001). *Trauma management: An emergency medicine approach*. St Louis, MO: Mosby Inc.

Fitzgerald, J. A. (2002). *Doctors and nurses working together: A mixed method study into the construction and changing of professional identities.* Unpublished Doctorate, University of Western Sydney, Sydney, NSW.

Fitzgerald, J., Lum, M., & Dadich, A. (2006). Scheduling unplanned surgery: A tool for improving dialogue about queue position on emergency theatre lists. *Australian Health Review, 30*(2), 219-231.

Fitzgerald, J. A., Lum, M., & Kippist, L. (2004, September 25-27). *Operating theatre bottlenecks: How are decisions about emergency theatre schedules made?* Paper presented at the 5th International CINet conference, Sydney, NSW.

Freidson, E. (1970). *Professional dominance: The social structure of medical care*. Hawthorne, NY: Aldine Publishing Company.

Gabel, R. A., Kulli, J. C., Lee, B. S., Spratt, D. G., & Ward, D. S. (1999). *Operating room management*. Woburn, MA: Butterworth-Heinemann.

Gordon, T., Paul, S., Lyles, A., & Fountain, J. (1998). Surgical unit time utilisation review: Resource utilisation and management implications. *Journal of Medical Systems, 12*, 169-179.

Hadley, R., & Forster, D. (1993). *Doctors as managers: Experiences in the front-line of the NHS*. Harlow: Longman.

Illich, I. (1977). Disabling professions. In I. Illich, I. K. Zola, J. McKnight, J. Caplan & H. Shaiken (Eds.), *Disabling professions* (pp. 11-39). London, England: Marion Boyars Publishers.

Jestin, P., Nilsson, J., Heurgren, M., Påhlman, L., Glimelius, B., & Gunnarsson, U. (2005). Emergency surgery for colonic cancer in a defined population. *British Journal of Surgery, 92*, 94-100.

Johnson, A. G. (2000). *The Blackwell dictionary of sociology: A user's guide to sociological language*. Oxford, UK: Blackwell Publishers.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*, 262-291.

Kim, S.-C., & Horowitz, I. (2002). Scheduling hospital services: The efficacy of elective-surgery quotas. *Omega, 30*(5), 335-346.

King, D., Kerridge, R., & Cansdell, D. (2004). *Ryde/Royal North Shore Hospital review*. Sydney, NSW: MA International.

Lebowitz, P. (2003). Why can't my procedures start on time? *Association of Operating Room Nurses Journal, 77*(3), 594-597.

Lum, M, & Fitzgerald, A. (2007). Dialogues for negotiating priorities in unplanned emergency surgical queues. In R. Iedema (Ed.), *Discourses of hospital communication: Tracing complexities in contemporary hospital organizations*. Basingstoke: Palgrave-Macmillan.

Magerlein, J. M., & Martin, J. B. (1978). Surgical demand scheduling: A review. *Health Services Research, 13*, 418-433.

Martin, D., & Singer, P. (2003). Canada. In C. Ham & G. Robert (Eds.), *Reasonable rationing: International experience of priority setting in health care* (pp. 42-63). Berkshire, England: Open University Press.

Myers, M. D. (1997). Qualitative research in information systems. *MIS Quarterly, 21*, 241-242.

NSW Health. (1998). *Booked patient and waiting time management: Operating guidelines*. Sydney, Australia: NSW Health.

NSW Health. (2002a). *NSW Health operating theatre management project report*. Sydney, Australia: NSW Health.

NSW Health. (2002b). *NSW Health operating theatre management project report: Supplementary report on information management survey*. Sydney, Australia: NSW Health.

Olesen, H. S. (2001). Professional identity as learning processes in life histories. *Journal of Workplace Learning, 13*(7/8), 290-297.

Parmigiani, G. (2002). *Modeling in medical decision making: A Bayesian approach*. West Sussex, England: John Wiley & Sons.

Pollicino, C., Haywood, P., & Hall, J. (2002). *Economic evaluation of the proposed surgical scheme at Auburn Hospital* (No. 19). Sydney, Australia: CHERE (Centre for Health Economics Research and Evaluation).

Redelmeier, D. A., Rozin, P., & Kahneman, D. (1993). Understanding patients' decisions: Cognitive and emotional perspectives. *Journal of the American Medical Association, 270*, 72-76.

Stake, R. E. (1995). *The art of case study research*. Thousand Oaks, CA: Sage Publications.

Western Canada Waiting List Project. (2001). *From chaos to order: Making sense of waiting lists in Canada: Final report*. Edmonton: WCWL Project, University of Alberta.

Wilson, T. (2005). *Structure in research methods.* Retrieved February 5, 2008, from http://informationr.net/tdw/publ/ppt/ResMethods/sld001.htm

Wright, I. H., Kooperberg, C., Bonar, B. A., & Bashein, G. (1996). Statistical modeling to predict elective surgery time: Comparison with a computer scheduling system and surgeon provided estimates. *Anaesthesiology, 85*, 1235-1245.

KEY TERMS

Clinical Determinants: Physical or physiological patient characteristics that influence the priority assigned to a surgical case, and in turn, influence patient health and economic outcomes. These include (but are not limited to) degree of risk to life, limb or senses, type of wound or injury, degree of haemodynamic stability, extent of pain, patient age, and diagnostic procedure required.

Emergency Surgery: Emergency surgery is required by a patient whose poor health requires urgent attention; this is most evident when there is risk to life, limb, or organ. An *emergency* is commonly understood to be an unforeseen combination of circumstances that demands immediate action, or an urgent need for assistance or relief. In the medical setting, it implies an injury or illness that poses an immediate threat to a person's health or life. In the clinical context, however, determining the immediacy and urgency of interventions becomes the ground for discourse between health care providers within the hospital setting. Each stakeholder offers different understandings of the term, *emergency* (Lum & Fitzgerald, 2007).

Human Technology: The body of information, skills, and experience developed for the production and use of goods and services.

Logistical Determinants: Practical or operational factors that influence the priority assigned to a surgical case, and in turn, influence patient health and economic outcomes. These include (but are not limited to) the availability of beds, hospital equipment, specialist instruments, surgery schedules, theatre time and space, as well as expert and support staff.

Non-Elective Surgery: See *emergency surgery*.

Professional Identity: An ongoing concern of the professional involving the practice of his or her work, social interactions with colleagues and patients, and his or her place within the professional institution and the professional discourse. It is thus socially bestowed, socially sustained, and socially transformed.

Scheduling: "A complex activity where human schedulers tend to make use of a wide range of knowledge, heuristics, and intuition" (Bharadwaj, Sen, & Vinze, 1999, p. 322).

Unplanned Surgery: Any surgical procedure that is not scheduled and thus, unexpected. While typically referred to as *emergency surgery*, the term *unplanned surgery* is preferred because of the variable understandings associated with the term *emergency surgery*.

I

Impact of RFID Technology on Health Care Organizations

Véronique Nabelsi
École Polytechnique de Montréal, Canada

Florina Stefanescu
ePoly Centre of Expertise in Electronic Commerce, Canada

INTRODUCTION

Radio Frequency Identification (RFID) technology has been considered the "next revolution in supply chain management" (Srivastava, 2004, p. 60). Current research and development related to RFID focuses on the manufacturing and retail sectors with the aim of improving supply chain efficiency. After the manufacturing and retail sectors, health care is considered to be the next sector for RFID (Ericson, 2004). RFID technology's potential to improve asset management in the health sector is considerable, especially with respect to asset management optimization.

In fact, health expenses have increased substantially in Organisation for Economic Co-operation and Development (OECD) countries in recent years. In Canada, the public health budget amounted to $91.4 billion (CAD) for the year 2005–2006 compared to $79.9 billion in 2003–2004 (CIHI, 2005). Moreover, the health care industry has been the focus of intense public policy attention. In order to curb this upward trend, the public heath sector in Canada is subject to strict budget constraints. Among the different alternatives for reducing expenditures, the improvement of asset management within the different health institutions appears to be worthwhile. RFID technology seems to be a viable alternative to help hospitals effectively manage and locate medical equipment and other assets, track files, capture charges, detect and deter counterfeit products, and maintain and manage materials. In other words, health care organizations would benefit particularly from RFID applications.

The main objective of this study is to investigate the potential for RFID technology within one specific supply chain in the health care sector. Based on a field study conducted in a large nonprofit hospital, this article tests some scenarios for integrating RFID technology in the context of two warehousing activities.

We will first introduce the context of the health care sector and the current applications of RFID technology in that sector. The next section presents the methodological approach that was used in the study. The research findings and their implications are then discussed. Finally, some closing remarks are made.

CURRENT CONTEXT OF THE HEALTH CARE SECTOR

The health care sector has been investing ever more money in information technology (IT) to reduce operating costs and improve patient safety and medical services, and RFID is expected to become critical to health care organizations in achieving these goals.

The IT implementation trend in health care works toward common IT platforms, which allow patient and product information to be exchanged. For many observers, the adoption and use of IT-related technologies, especially RFID, by health care organizations could boost the effectiveness and efficiency of this information-intensive sector. However, the health care sector has been relatively slow to embrace the full potential of IT initiatives. In general, the implementation of IT in hospitals has not been particularly successful (Aarts, Doorewaard, & Berg, 2004; Hersh, 2004; Pare, 2002). The major impediments appear to be linked more to organizational issues than to technological problems (Southon, Sauer, & Dampney, 1997). Among the many factors slowing down the implementation of IT initiatives, previous studies have identified the complexity of health care organizations (Glouberman & Mintzberg, 2001; Glouberman & Zimmerman, 2002), their inappropriate organizational structure (Mintzberg, 1979), and integration issues (Christensen, Bohmer, & Kenagy, 2000; Kumar, Subramanian, & Strandholm, 2002).

Characteristics of the Health Care Supply Chain

In response to all those constraints, some visionaries understand and are already taking action to rectify supply chain processes as a key strategic factor supporting patient service. A study within certain hospital departments made it possible to identify the priorities (Landry & Beaulieu, 1999). The priorities of administrative departments are generally the review and improvement of the supply chain process, IT system modernization, and system integration on a common platform used by other health care organizations.

In fact, Beaulieu, Duhamel, and Martin (2004) state that the integration of procurement activities would improve efficiency by eliminating nonvalue-added activities; this would allow health care organizations to concentrate on more strategic activities (Landry & Beaulieu, 1999). According to some researchers, better resource monitoring and allocation will reduce costs throughout the restocking chain (Blouin, Beaulieu, & Landry, 2001; Perrin, 1994). In addition, the procurement activities represent a large proportion of health expenses. In a hospital, for instance, they range from 30% to 46% of all expenses (Bourgeon et al., 2001; Poulin, 2004). Moreover, the health sector currently loses up to 15% of its assets due to inappropriate and inefficient monitoring procedures (Nabelsi, 2007). The larger the hospital, the bigger these problems are (Nabelsi, 2007).

RFID APPLICATIONS IN HEALTH CARE ORGANIZATIONS

The market potential is interesting: according to a recent study, the worldwide market for RFID tags (active, passive and semi-active) and systems in the health sector will rise from $90 million in 2006 to $2.1 billion in 2016 (Harrop & Das, 2006). The sale and implementation of a complete RFID platform solution would obviously represent much higher revenues for the participating solution providers.

The application of RFID in hospitals has received a great deal of attention over the last few years, with a "boom" in early 2003 due to the rapid spread of Severe Acute Respiratory Syndrome (SARS) in Taiwan (Li, Liu, Chen, Wu, Huang, & Chen, 2004; Wang, Chen, Ong, Liu, & Chuang, 2006). This emergent technology plays a significant role in the global fight to contain SARS; some hospitals have tested an RFID infrastructure system that tracks the movement of patients, medical professionals, and visitors in order to trace and identify when and where people may have been in contact with patients infected by SARS (Li et al., 2004; Wang et al., 2006). The real-time information can be received from this system to trace detailed patient medical records and read bio-information such as pulse, temperature, and respiratory rate (Li et al., 2004).

A number of hospitals have announced new projects including several that will use RFID for supply chain management applications. Tracking assets has become a potential area for improving hospitals' performance (Anonymous, 2006a). In addition, the use of RFID technology for equipment tracking in the health care supply chain can lead to a tremendous reduction in inventory levels and better collaboration among supply chain players. For example, RFID can reduce the time staff members spend looking for equipment they need, thereby improving the utilization rate of equipment and cutting down on missing equipment (Ostbye et al., 2003). Furthermore, the greatest use of RFID in the health care sector will be for labels on medical equipment, drugs, and other products at the item level and the infrastructure and services to support this throughout the supply chain. It will also be used in health care facilities to protect products against counterfeiting. The primary purpose will be to prevent counterfeiting by establishing the full history of a given package at all times—known as its pedigree.

Furthermore, RFID technology can trace patients and materials in hospitals in order not only to optimize the health care process but to minimize human errors. The Food and Drug Administration (FDA) is paying increasing attention to RFID, especially to avoid mistakes because errors involving medication and drug administration are very expensive for the health care sector. A number of studies state that about 7% of patients (Bates, Cullen, & Laird, 1995) in the USA experience an adverse drug experience (ADE) and that the costs of these amount to about US$80 billion annually (Johnson & Bootman, 1995). In addition, clinical studies suggest that Bar Code Medication Administration (BCMA) might significantly decrease ADEs by up to 58% (Jensen, Merry, Webster, Weller, & Larsson, 2004). Another use that has been suggested is cross-checking of blood transfusions (Anonymous, 2006b). A pilot study in Massachusetts General Hospital used the technology to reduce blood transfusion errors by using RFID tags in patient wristbands (Anonymous, 2005). As in many studies, patients were fitted with RFID bracelets containing patient information, including blood type, allergies, and more. Overall, RFID initiatives in the health care sector are driven by the desire to enhance patient safety and improve quality assurance.

METHODOLOGY

Our study builds on previous work (Strassner & Schoch, 2004; Subirana et al., 2003) and focuses on a large nonprofit hospital supply chain which is considering implementing RFID technology.

Research Design

Since the main objective of this case study is to gain a better understanding of the potential for RFID technology in the context of warehousing activities in one specific supply chain, the research design corresponds to an exploratory research initiative. This appears appropriate since it enables researchers to examine a new interest. Moreover, a "case study is a research strategy which focuses on understanding the dynamics present within single settings" (Eisenhardt, 1989, p. 534). This research strategy is consistent with the goal of our study. The field research was conducted in 12 consecutives steps. This methodology was developed at the ePoly research center (see Figure 1, adapted from Lefebvre, Lefebvre, Bendavid, Fosso Wamba, & Boeck, 2005).

The first six steps correspond to an initial phase which can be broadly termed as the "opportunity-seeking phase."

- **Step 1:** All health care organizations are highly motivated to find the best alternatives to give patients better care and reduce time-consuming human administrative activities.
- **Steps 2 and 3:** Benefits are observed in several areas of opportunity related to product information which is to be exchanged with partners and/or departments within the same organization.
- **Steps 4 and 5:** Reflect the current context in terms of supply chain structures and dynamics, and existing intra- and inter-organizational business processes. The implementation of integrated IT platforms allows systems to be integrated and facilitates the exchange of information.

The second phase constitutes "scenario building" to evaluate RFID opportunities; this phase incorporates steps 6 to 10.

- **Steps 6 and 7:** Several scenarios are evaluated regarding the implementation of RFID in applications such as tracking patients, equipment, and errors.
- **Step 8:** Business and technological concerns are evaluated. RFID implementation provides increased integration with suppliers as information is captured in the RFID tags. This minimizes documentation exchanges, dramatically improves the quality of information transferred and reduces the need for internal control or manual activities to zero.
- **Step 9:** Business processes are redesigned to integrate RFID technology. By including enterprise resource planning (ERP) and middleware integration, steps are taken toward process automation; some of those steps have already been accomplished in some health care organizations. A free flow of information, the identification of impacts on human resources and supply chain alignment are assets in redesigning and simplifying the business process activities.
- **Step 10:** Several scenarios are simulated. Suggestions are received from supply chain departments, regarding replenishment needs in the ER, product validation at the time of receipt, tracking mobile beds and chairs, and so forth.

The final phase is to "validate the scenarios" retained from the second phase in a controlled environment (Proof of Concept or PoC; step 11) and then in a real-life setting (step 12).

- **Step 11:** PoC in laboratory. Product receiving and/or replenishment are reproduced in a similar physical and technological environment to that found in the health care organization. The main goal is to demonstrate the feasibility of RFID technology and assess the ERP and middleware integration, process automation, information flow, and human resources impact for all supply chain members. The equipment is acquired, calibrated, and configured, and the business rules are identified and configured in various middleware applications and integrated with the ERP engine. Finally, dry-run tests are conducted to validate

Figure 1. Steps undertaken in the field study

Detailed activities	
Phase 1: Intra- and inter-firm opportunity seeking	
Step 1	**Determination of the primary motivation** towards the use of RFID and the EPC network **(WHY?)**
Step 2	**Analysis of the product value chain** (PVC) activities specific to a given product **(WHAT?)**
Step 3	**Identification of the critical activities in the PVC**, (**WHICH** activities to select and **WHY?)**
Step 4	**Mapping of the network of firms supporting the PVC** to understand the relationships between the firms supporting the product **(WHO** and **WITH WHOM?)**
Step 5	**Mapping of intra- and inter-business processes** for critical activities **(«As is»)** (**HOW** within and between organisations?).
	Time and motion data capture and analysis
Phase 2: Intra- and inter-firm scenario building integrating RFID and the EPC network	
Step 6	**Evaluation of RFID and the EPC network opportunities** with respect to the product (level of granularity), to the firms involved in the SC and to the specific PVC activities
Step 7	**Evaluation of RFID and the EPC network potential applications** including scenario building and process optimization («As could be») (**HOW** within and between organisations?)
Step 8	**Mapping of intra- and inter-organizational processes integrating RFID and the EPC network**
Step 9	**Validating business processes and technological solutions integrating RFID and the EPC network** with key respondents. Feasibility analysis and evaluation of the challenges **including ERP and middleware integration**, **process automation, information flow, huma**
Step 10	**Simulating several scenarios** for final choice of proof of concept
Phase 3: Scenario validation and demonstration	
Step 11	**Proof of Concept (PoC) in laboratory** simulating physical and technological environments, and, interfaces between SC players. Feasibility demonstration of RFID technology and evaluation assessment (ERP and middleware integration, process automation, inform a. Laboratory set-up -Equipment acquisition -Equipment set-up, calibration and configuration -Business rules define and configure in various Middleware -IS integration (e.g. Middleware with ERP) -Dry-run to test IT integration and process integration at all the SC members' level -Business rule refine and IT adjustment b. PoC, post-analysis and decision to go for the beta test replicating PoC scenarios in a real-life setting
Step 12	**Beta Test in real life setting.** Deployment of application and its appropriation by the different organizations involved and their staff

the IT and process integration at the level of all supply chain members.

- **Step 12:** Based on the analysis of PoC results, the decision to run beta tests in a real-life setting is made; the application is deployed in the health care organization and appropriated by the staff involved.

Research Sites

One hospital was involved in the research design. This hospital is briefly described in the following paragraph.

Hospital Profile

The nonprofit hospital is a regional health care facility and is one of the major hospitals affiliated with the Université de Montréal; it is one of the largest hospitals in Quebec (North America). The hospital, which employs 4,000 persons currently, has 554 beds. It has an operating budget totaling $231 million and $66 million of charges associated with supply chain activities. The hospital is a university center that provides specialized and ultraspecialized services to a regional and supraregional clientele. The hospital uses the systems, applications, products (SAP) platform to integrate and centralize patients' personal data such as patient indices from different sources, scheduling and visit information, patient movements (admission, transfer, and discharge information) and gather activity volumes related to clinical data from various sources, all of which are valuable for the management of finances and controls. The supply chain activity is a target for integration with external suppliers using the global healthcare exchange (GHX) platform through an e-supply chain, by automating the ordering process and electronic invoices. Probably the most complex IT integration process will start with the exchange of a single "Electronic Patient File" with health care organizations all over Canada.

Data Collection

The case study methodology allowed us to employ multiple data collection methods to gather information from one or a few entities (Benbasat, Goldstein, & Mead, 1987). Data collection for the case study approach was based on:

1. Various focus groups were conducted in the university-based research center with health care professionals (5) and IT experts (3). The main objective of these focus groups was to reach a consensus on strategic intent with respect to the use of RFID technology in one product value chain (steps 1, 2, and 3), and to evaluate different scenarios and retain the "preferred" or "as could be" scenario (steps 7, 8, and 9). Each consecutive step of the methodology illustrated in Figure 1 was evaluated and validated and agreed upon with members of the focus groups.
2. On-site observations at the research site were performed in order to carry out the process mapping required for Steps 5, 6, and 9. The analysis of the current inter-organizational business processes allows the researchers to understand the supply chain dynamic and its business environment.
3. Semistructured interviews were conducted with seven participants at the research site. The participants in the case study were the department head responsible for the logistics and distribution division (1) and some warehousemen (6). The purpose was to document and obtain more detailed information, resolve potential inaccuracies in the mapping of existing business processes, and ensure that our observations and the results of the mapping were valid and representative of the normal flow of operations (Steps 5 and 6).

The researchers acted as observers, interviewers, and facilitators (for focus groups). They also formulated and presented the detailed scenarios that were developed from the empirical evidence gathered in the nonprofit hospital.

RESULTS AND DISCUSSION

The health care professionals involved in this study are aware of the potential advantages derived from the implementation of RFID technology. Their initial motivations were focused on the reduction of manual interventions by employees for the overall "receiving" and "verifying" processes and the increased accuracy of validation of controls such as validate delivery order (DO) against purchase order (PO), and quantities of items received vs. ordered.

Within the scope of this article, our discussion will mainly focus on the empirical results obtained from steps 5, 6, and 9 (Figure 2). These three steps correspond to the mapping of current business processes and redesigned processes integrating RFID technology. Proponents of the process-based approach believe that the process view is "a more dynamic description of how an organization acts" (Magal, Feng, & Essex, 2001, pp. 2). Moreover, this process view provides a structured approach which allows one to focus on value creation. The process view is also used increasingly often to evaluate the impact of information technologies (Subramaniam & Shaw, 2004).

Current Business Processes

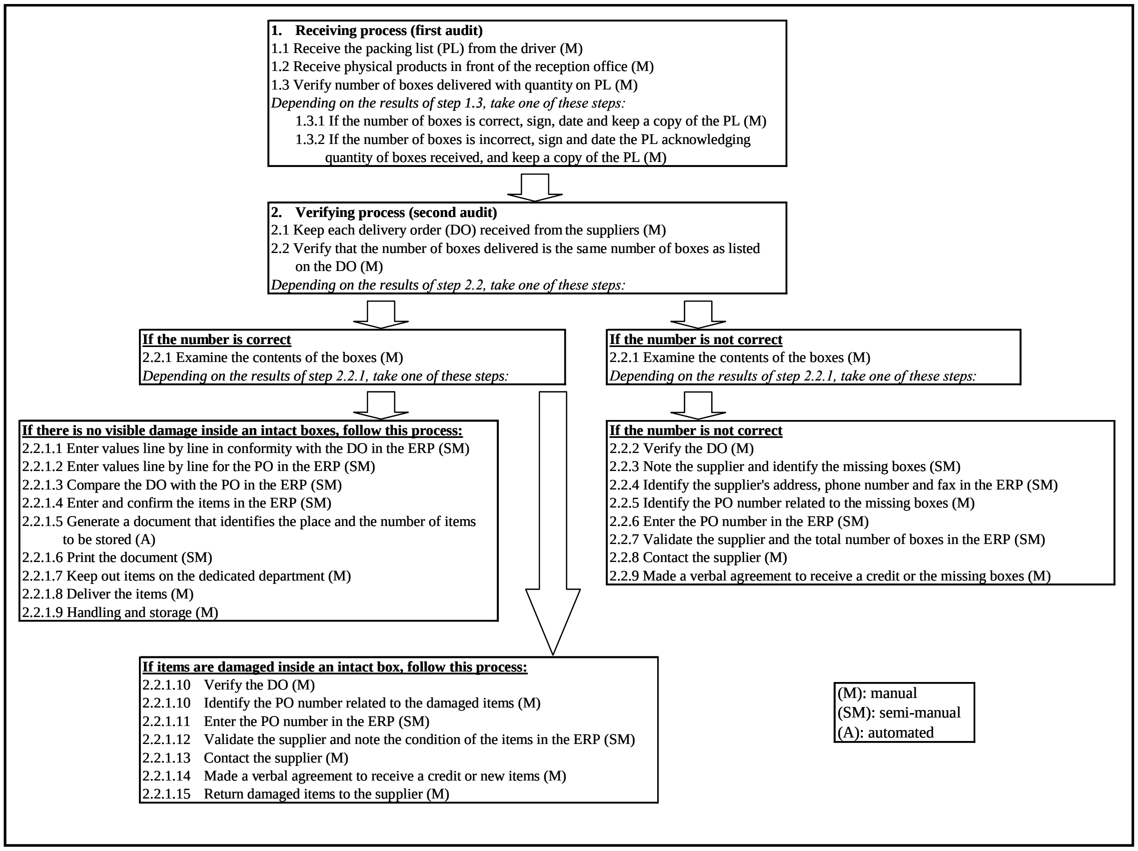
In Figure 2, processes are drilled down from the more general to the more detailed. Based on the analysis of the current inter-organizational processes in that figure, the following observations can be made: (a) the overall "receiving" and "verifying" processes require time-consuming human interventions by the warehousemen (they depend on manual resources); (b) time is devoted to several audit levels (receiving and verifying); (c) validation is error-prone due to of the numerous visual checks and repetitive manual activities when the DO is compared line by line against the packing slip (PS).

Based on the analysis of these processes, the process flow of receiving and verifying contains several steps to be completed in a semi-manual verification of 15 to 17 steps, depending on the following situations:

a) The best case scenario reveals no anomalies in the receiving and verification processes, which mean that the products and their quantities are as ordered. However, the audit processes require manual verification of all 17 processes: 11 manually and 5 semimanually. Only one is automatically created by SAP (2.2.1.5), a document that identifies the place and the number of items to be stored. We can observe that even in this best-case situation, all those steps need to be performed in order to meet this expectation.

b) If the second audit reveals inadequate quantities inside boxes or products are damaged for any reason, 15 or 16 other manual or semimanual steps must be completed to expose those anomalies.

Figure 2. Current interorganizational processes



In general, all receiving and verifying steps seem to fall into two categories (manual and semimanual).

Retained Scenario Integrating RFID Technology

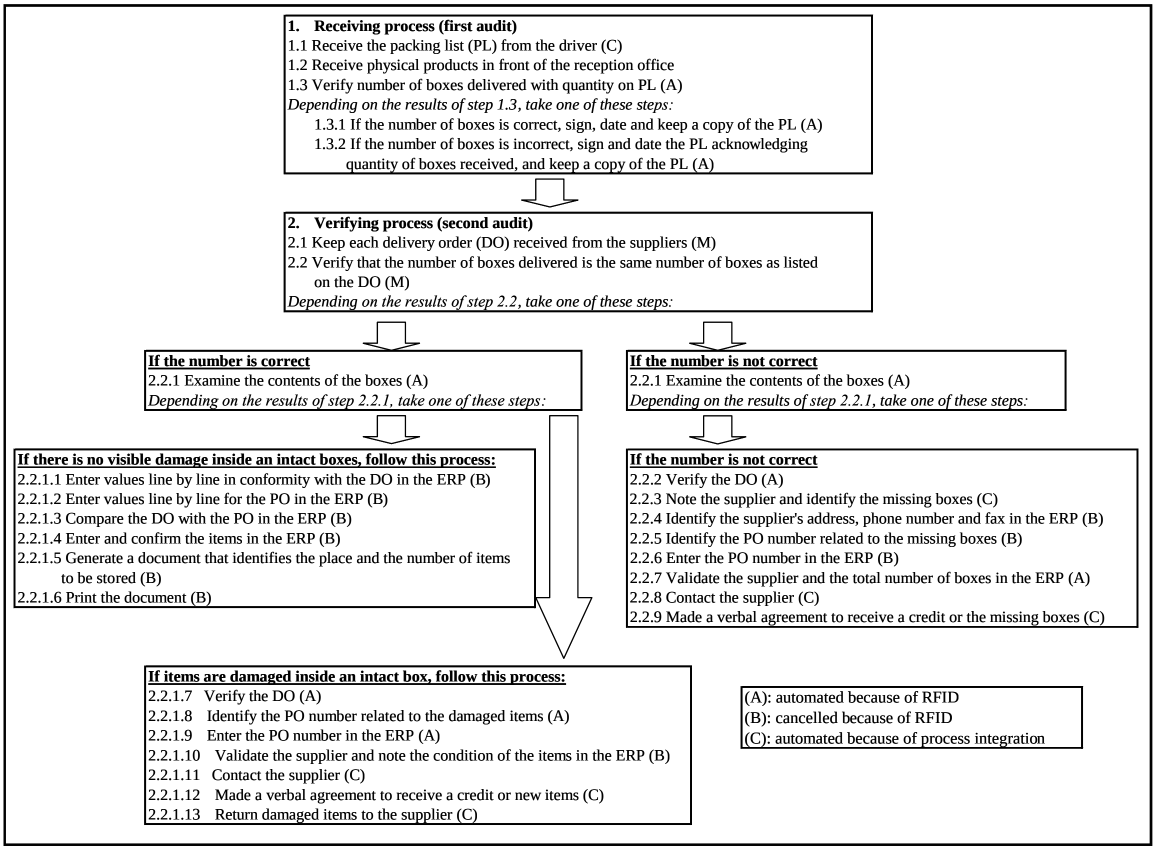
The retained scenario integrating RFID technology (steps 8 and 9) was validated with the focus groups in a nonprofit hospital scenario. By contrasting the existing business processes (Figure 2) with those integrating RFID technology (Figure 3), it is possible to explore some of the impacts of RFID technology on the receiving and verifying processes.

When one integrates RFID (tags on boxes or, even better, tags on products inside boxes) new categories of business process emerge.

Automated processes will reduce human interventions, and therefore will protect against human error and assure consistency of data. Some of the manual processes described in Figure 1 will be eliminated because of automation, others reduced because processes are integrated with the SAP supply chain. In the first situation in the described current process, the number of observed processes will be reduced by 14 (six processes will be automated, six will be eliminated, and two will be reduced because of automated integration with the existing SAP system).

For the second situation, a substantial number of steps will be automated either because of RFID or because of the integration of the RFID and SAP systems (seven processes) or will no longer be necessary (two processes eliminated). All information-based processes are now automatically performed at the receiving point, providing a high level of accuracy of information at a high level of granularity. With RFID technology, approximately 80% of the processes will therefore be replaced or automated, resulting in a proportional human intervention time.

Figure 3. The impact of RFID on selected business processes



FUTURE TRENDS

Increased operating profits and greater customer satisfaction are the objectives that health care organizations strive for. The main objectives are better tracking in real time of patients' presence and patient records, tracking patients in case of contagion, avoiding medical errors such as patient misidentification, reducing medical errors by increasing the accuracy of reads, assuring better arrangement and distribution of products especially for emergency interventions, avoiding confusion during emergencies by quickly tracking the availability and location of a specific blood group, tracking equipment to avoid theft of equipment and supplies, and combating the growth of counterfeit medicine. Reverse logistics and recall management (returning shipments because they are out of date or stocking conditions are inadequate) are new requirements for increased supply chain cost efficiencies.

For health care organizations, RFID is the inevitable next step towards the new generation of health care service operations. It should provide new efficiencies, improved services, and greater productivity by reducing the time health care professionals spend in locating misplaced equipment. It will also allow cost avoidance by preventing spending on redundant equipment, enhance health care workflows and increase patient care in organizations seeking a competitive advantage.

CONCLUSION

This article assesses the impacts of RFID technology on information flows within one health care supply chain in a large nonprofit hospital. Our study shows that RFID technology could enhance information flows within a supply chain by automating all information-based activities, reducing inventory and warehousing costs, reducing human errors, and managing real-time product information throughout the supply chain.

The synchronization of information and material flows in a given supply chain is the principle upon which supply chain integration is based. However, in the health care sector, supply chain integration cannot rely on traditional logistical information to become more efficient. In health care delivery, the unpredictable fluctuations in the demand for services, but also the need to meet shorter lead times and very high service levels, are leading health care organizations to develop volume-flexible strategies (Jack & Powers, 2004). RFID is one technology that has been proven to provide benefits in the hospital environment.

REFERENCES

Aarts, J., Doorewaard, H., & Berg, M. (2004). Understanding implementation: The case of a computerized physician order entry system in a large Dutch university medical center. *Journal of the American Medical Informatics Association*, *11*(3), 207-216.

Anonymous. (2005). RFID knowledgebase - Massachusetts General Hospital blood. Retrieved February 5, 2008, from http://www.idtechex.com/knowledge/en/casestudy.asp?freefromsection=5

Anonymous. (2006a). *Exavera technologies – e-shepherd overview*. Retrieved February 5, 2008, from http://www.exavera.com/healthcare/eshepherd.php

Anonymous. (2006b). *Patient safety applications of bar code and RFID*. Zebra Technologies White paper.

Bates, D., Cullen, D., & Laird, N. (1995). Incidence of adverse drug events and potential adverse drug event. *Journal of the American Medical Association*, *274*, 29-34.

Beaulieu, M., Duhamel, C., & Martin, R. (2004). Intégrer les considérations logistiques au réaménagement d'un bloc opératoire [Special edition]. *Logistique et Management*, 93-99.

Benbasat, I., Goldstein, D.K., & Mead, M. (1987). The case research strategy in studies of information systems. *MIS Quarterly*, *11*(3), 369-386.

Blouin, J.P., Beaulieu, M., & Landry, S. (2001). Systèmes de réapprovisionnement des unités de soins: description et implications organisationnelles. *Logistique & Management*, *9*(2), 43-53.

Bourgeon, B., Constatin, A., Karalszyk, J.F, Pedrini, S., Landry, S., Diaz, A., et al. (2001). Évaluation des coûts de logistiques hospitalières en France et aux Pays-Bas, *Logistique & Management*, *9*(1), 81-87.

Canadian Institute for Health Information (CIHI). (2005). *Preliminary provincial and territorial government health expenditure estimates, 1974–1975 to 2006–2007.*

Christensen, C.M., Bohmer, R., & Kenagy, J. (2000). Disruptive innovations cure health care? *Harvard Business Review*, *78*(5), 102-120.

Eisenhardt, K.M. (1989). Building theories from case study research. *Academy of Management*, *14*(4), 532-550.

Ericson, J. (2004). RFID for hospital care. Line 56. Retrieved February 5, 2008, from http://www.line56.com

Glouberman, S., & Mintzberg, H. (2001). Managing the care of health and the cure of disease – part I : Differentiation. *Health Care Management Review*, *26*(1), 56-69.

Glouberman, S., & Zimmerman, B. (2002). Complicated and complex systems: What would successful reform of medicare look like? *Commission on the Future of Health Care in Canada*.

Harrop, P., & Das, R. (2006). *RFID in healthcare 2006-2016*. IDTechEx.

Hersh, W. (2004). Editorials: Health care information technology progress and barriers. *Journal of American Medical Association*, *292*(18), 2273-2274.

Jack, E.P., & Powers, T.L. (2004). Volume flexible strategies in health services: A research framework. *Production and Operations Management*, *13*(3), 230-244.

Jensen, L.S., Merry, A., Webster, C., Weller, J., & Larsson, L. (2004). Evidence-based strategies for preventing drug administration errors during anaesthesia. *Anaesthesia*, *59*, 493-504.

Johnson, J., & Bootman, J. (1995). Drug-related morbidity and mortality: A cost-of-illness model. *Archives of Internal Medicine*, *155*, 949-1956.

Kumar, K., Subramanian, R., & Strandholm, K. (2002). Market and efficiency-based strategic responses to environmental changes in the health care industry. *Health Care Management Review*, *27*(3), 21-31.

Landry, S., & Beaulieu, M. (1999). L'approvisionnement face aux bouleversements du secteur de la santé: le cas québécois. *Revue Internationale de l'achat*, *19*(3/4), 17-25.

Lefebvre, L.A., Lefebvre, É., Bendavid, Y., Fosso Wamba, S., & Boeck, H. (2005). The potential of RFID in warehousing activities in a retail industry supply chain. *Journal of Chain and Network Science*, *5*(2), 101-111.

Li, C.J., Liu, L., Chen, S.Z., Wu, C.C., Huang, C.H., & Chen, X.M. (2004, March 21-23). Mobile healthcare service system using RFID. In *Proceedings of the 2004 IEEE International Conference on Networking, Sensing & Control Taipei*, Taiwan, (pp. 1014-1019).

Magal, S.R., Feng, M., & Essex, P.A. (2001). An exploratory study of Web-based electronic commerce applications. *Journal of Information Technology Theory and Application*, *3*(5), 139-149.

Mintzberg, H. (Ed.). (1979). *The structuring of organizations: A synthesis of the research*. Englewood Cliffs, NJ: Prentice-Hall.

Nabelsi, V. (2007). *Processus de réapprovisionnement dans le secteur de la santé: collaboration et performance*. Unpublished doctoral dissertation, École Polytechnique de Montréal, Dissertation, Montréal.

Ostbye, T., Lobach, D., Cheesborough, D., Lee, A., Krause, K., Hasselblad, V., et al. (2003). Evaluation of an infrared/radiofrequency equipment-tracking system in a tertiary care hospital. *Journal of Medical System*, *27*(4), 367-380.

Pare, G. (2002). Implementing clinical information systems: A multiple-case study within a US hospital. *Health Services Management Research*, *15*(2), 71-92.

Perrin, R.A. (1994). Exchange cart and par level supply distribution systems: Form follows function. *Hospital Materiel Management Quarterly*, *15*(3), 63-76.

Poulin, E. (2004). *L'analyse comparative du processus de logistique hospitalière : un remède aux maux du secteur de la santé ?*. CMA Management, Dernier numéro. Retrieved February 5, 2008, from http://www.managementmag.com3index.cfm/ci_id/1671/la_id/2

Southon, F.C.G., Sauer, C., & Dampney, C.N.G.K. (1997). Information technology in complex health services: Organizational impediments to successful technology transfer and diffusion. *Journal of the Americam Medical Informatics Association*, *4*(2), 114-124.

Srivastava, B. (2004). Radio frequency ID technology: The next revolution in SCM. *Business Horizons*, *47*(6), 60-68.

Strassner, M., & Schoch, T. (2004). *Today's impact of ubiquitous computing on business.* Institute of Information Management of University of St. Gallen. Retrieved February 5, 2008, from www.vs.inf.ethz.ch/publ/papers/Strassner-Schoch_Impact-Ubicomp.pdf

Subirana, B., Eckes, C., Herman, G., Sarma, S., & Barrett, M. (2003). *Measuring the impact of information technology on value and productivity using a process-based approach: The case for RFID technologies*. Retrieved February 5, 2004, from http://ssrn.com/abstract_id=478582

Subramaniam, C., & Shaw, T M.J. (2004). The effects of process characteristics on the value of B2B e-procurement. *Information Technology and Management*, *5(*1/2), 161-180.

Wang, S.W., Chen, W.H., Ong C.S., Liu, L., & Chuang, Y.W. (2006). RFID applications in hospitals: A case study on a demonstration RFID project in a Taiwan hospital. In *Proceedings of the 39th Hawaii International Conference on System Sciences* (pp. 1-10).

KEY TERMS

Active RFID Tags: Active tags contain a battery that powers the microchip and allows it to transmit a signal to the reader.

Passive RFID Tags: Passive tags rely heavily on the magnetic field of radio waves to generate a current which can be received by their antenna.

Receiving Process: Handling products that arrive at the receiving office and acknowledging receipt and unpacking supplies.

RFID: RFID, which belongs to the large landscape of Automatic Identification and Data Capture technologies (AIDC), uses radio waves to automatically identify in real-time individual objects, items, or products. An RFID system is basically composed of three major layers: a tag containing a chip, a reader and its antennas, and a computer.

Semi-Active RFID Tags: Semi-active tags draw power from the magnetic field created for reader-to-reader communication.

Verifying Process: The function encompasses the physical receipt of material, the inspection of the shipment for conformance with the purchase order including quantity, quality and damage, the identification of the destination and delivery there, and the preparation of receiving reports.

Volume Flexibility: In health care, this represents a means to improve service delivery and it allows organizations to leverage their scarce resources for optimal utilization in response to fluctuations in patient demand.

Implementing RFID Technology in Hospital Environments

Marlyn Kemper Littman
Nova Southeastern University, USA

INTRODUCTION

A promising approach for facilitating cost containment and reducing the need for complex manual processes in the healthcare space, RFID (Radio Frequency Identification) technology enables data transport via radio waves to support the automatic detection, monitoring, and electronic tracking of objects ranging from physicians, nurses, patients, and clinical staff to walkers, wheelchairs, syringes, heart valves, laboratory samples, stents, intravenous pumps, catheters, test tubes, and surgical instruments (Karthikeyan & Nesterenko, 2005). RFID implementations streamline hospital applications and work in concert with WLANs (wireless local area networks) and mobile devices such as cellular phones and personal digital assistants (PDAs). RFID technology also safeguards the integrity of the drug supply by automatically tracing the movement of medications from the manufacturer to the hospital patient.

This article begins with a discussion of RFID development and RFID technical fundamentals. In the sections that follow, the work of standards organizations in the RFID space is introduced, and capabilities of RFID solutions in reducing costs and improving the quality of healthcare are described. Descriptions of RFID initiatives and security and privacy challenges associated with RFID initiatives, are explored. Finally, trends in the use of RFID-augmented wireless sensor networks (WSNs) in the healthcare sector are introduced.

BACKGROUND

RFID technology traces its origins to 1891 when Guglielmo Marconi first transmitted radio signals across the Atlantic Ocean, and demonstrated the potential of radio waves in facilitating data transport via the wireless telegraph. During the 1930s, Alexander Watson Watt discovered radar, and illustrated the use of radio waves in locating physical objects. Initially used in World War II in military aircraft in what is now called the first passive RFID system, radar technology enabled identification of incoming aircraft by sending out pulses of radio energy and detecting echoes (Want, 2004). Libraries have used RFID technology for electronic surveillance and theft control since the 1960s.

Present-day RFID solutions track objects ranging from tools at construction sites and airline baggage, to dental molds and dental implants. RFID systems monitor the temperature of perishable fruit, meat, and dairy products in transit, in order to ensure that these goods are safe for consumption, and facilitate the detection of package tampering and product recalls (Want, 2005). The U.S. Department of Defense (DoD) mandates the use of RFID tags as replacements for barcodes for tracking goods (Ho, Moh, Walker, Hamada, & Su, 2005), and requires suppliers to use RFID tags in equipment and clothing shipped to military personnel. RFID technology is widely used by major retailers that include Home Depot and Wal-Mart in the U.S., and Marks and Spencer in the United Kingdom to track inventory. In the transportation and education sectors, credit cards that incorporate RFID technology enable automatic transactions at gas stations and toll plazas and at university bookstores, libraries, and cafeterias. RFID systems also facilitate building access, port security, vehicle registration, and supply chain management; verification of the identity of pre-authorized vehicles and their drivers at security checkpoints; and reduction in the circulation of counterfeit goods and paper currency (Garfinkel, Jules, & Pappu, 2005).

Developed by the U.S. Department of Energy Oak Ridge National Laboratory (ORNL), the RFID-enabled Protected Asset Tracking System is the first instance of RFID technology installed at the National Nuclear Security Administration site (Oak Ridge National Laboratory, 2007). Capabilities of the ORNL-sponsored RFID accountability system (RAS) in accurately tracking assets and monitoring the location of personnel were validated in a pilot test conducted in a large-sized facility at the Washington Navy Yard. This initiative also demonstrated the role of RFID technology in facilitating the safety of rescue workers responding to crisis situations and national emergencies in federal facilities, and the importance of using RFID tags on miners, firefighters, and other workers in hazardous occupations, so their location and safety can be monitored as well.

RFID TECHNICAL FUNDAMENTALS

RFID systems consist of RFID tags or transponders, and interrogators or readers. Classified as passive, semiactive, or active, a RFID tag is an extremely small device containing a microchip, also called a silicon chip or integrated circuit that, at a minimum, holds digital data in the form of an EPC (Electronic Product Code). RFID tags are affixed to or incorporated into objects, such as persons or products (Weinstein, 2005).

A RFID tag is also equipped with an antenna for enabling automatic receipt of and response to a query from an RFID interrogator, via radio waves (Myung & Lee, 2006). The RFID communications process involves the exchange of an electromagnetic query and response, thereby eliminating RFID dependency on direct line-of-sight connections. Subsequent to transmission of the EPC from the RFID tag to the RFID interrogator, the tagged object can be monitored and traced.

Passive RFID tags are inexpensive and limited, in terms of functions supported (Weinstein, 2005). In terms of transmission, a passive nonbattery operated RFID tag makes use of incoming radio waves when it is within range of a RFID interrogator to transmit a response. A passive RFID tag contains the EPC in the form of eight-bit data strings associated with a distinct object and several bits of memory for storing data describing the tagged object. When multiple passive RFID tags transmit EPCs concurrently in response to RFID interrogators, collisions occur, thereby disrupting information flow. Designed to support passive RFID tag operations, the adaptive binary splitting (ABS) collision arbitration protocol diminishes the occurrence of collisions, thereby significantly reducing delay and communications overhead in the transmission process (Myung & Lee, 2006).

As with passive tags, semiactive and active RFID tags also feature EPCs or unique identifiers, and utilize the RF spectrum for data transmission. Batteries in semiactive RFID tags remain dormant until signals are received from interrogators. When sufficient power is available, semiactive tags initiate data transmissions in response to interrogator queries.

An active RFID tag features an onboard battery that serves as its own power source for performing operations, and transmitting the EPC and related data on-demand in response to interrogator queries. An active tag also supports security functions, and can contain an environmental sensor (Karygiannis, Eydt, Barber, Bunn, & Phillips, 2006). Moreover, an active tag enables transmissions over longer distances than passive and semiactive tags. However, an active tag is limited in sustaining continuous operations as a consequence of battery constraints. Since active tags are larger and more costly to implement than passive and semiactive tags, these tags typically monitor large items.

Classified as active RFID systems, wireless sensor networks (WSNs) contain sensors for monitoring the environment (Philipose, Smith, Jiang, Mamishev, Roy, & Sundara-Rajan, 2005). To conserve available power and provision efficient operations, active RFID tags that are part of WSNs are also equipped with battery-powered four-byte or eight-byte processors, and employ cryptographic algorithms to facilitate secure transmissions. Next-generation sensors are expected to operate without batteries by harvesting power from ambient sources.

RFID operations are typically carried out in the unlicensed portion of the RF spectrum, and are dependent on the availability of suitable frequencies and bandwidth at affordable costs. Generally, active RFID tags operate in allocated spectrum in the UHF (ultra high frequency range) between the 300 MHz (megahertz) and 3000 MHz frequencies, and the SHF (super high frequency) between the 3 GHz (gigahertz) and 30 GHz frequencies. These tags carry more data, and support operations over longer distances than passive RFID tags operating in allocated spectrum in the LF (low frequency) between the 30 kHz (kilohertz) and 300 kHz frequencies, and the HF (high frequency) between 3 MHz and 30 MHz frequencies (Littman, 2002). It is important to note that while increased transmission rates and operating range are advantageous in the UHF and SHF spectrum, data transmitted via RFID systems in these frequencies are also vulnerable to interception. Spectrum allocation for RFID utilization differs between countries and regions, and between member states in the European Union.

RFID SPECIFICATIONS

Developed by EPCglobal, the Electronic Product Code (EPC) replaced barcodes developed during the 1970s. An EPC is a unique number that corresponds to an individual product or a container of products. EPCs are embedded in RFID tags that are affixed to or incorporated into an object. As a consequence, the terms EPC tag and RFID tag are used interchangeably. With an EPC and access to the EPC network, detailed information about a product, item, or package, such as dates of manufacture and expiration, can be determined. Designed for companies using RFID technology for supply chain management, the EPC network provisions real-time access to RFID data via the Internet. The EPC network also enables product authentication and exchange of product information among authenticated participants (Staake, Thiesse, & Fleisch, 2005).

EPCglobal is a participant in GS1, an organization with members in more than 100 countries that supports the design and implementation of global technologies and supply chain standards. The GSI HUG (Healthcare User Group) supports development of automatic product identification specifications for EPC-compliant RFID devices to enhance patient safety and facilitate efficiencies in supply chain management. In addition to EPCglobal, entities active in establishing RFID standards include the European Telecommunications Standards Institute (ETSI) and the International Telecommunications Union—Telecommunications Standards Sector (ITU-T).

RFID INITIATIVES IN THE HEALTHCARE SPACE

RFID initiatives in the healthcare sector vary in scope and complexity, and support a broad array of applications. For instance, RFID systems consisting of ultra-low power RFID tags monitor patients with medical implantations such as pacemakers, insulin pumps, and defibrillators. The medical identifier embedded in the medical implantation also provides access to the medical history and relevant data including the name of nonverbal or unconscious patients. Hospitals increasingly mandate the use of wristbands embedded with RFID tags for verification of a patient's identity. RFID tags for surgical patients contain additional data including the type of procedure and the name of the surgeon to ensure optimal outcomes.

RFID systems reduce medical errors including incidents of drug over dosage by ensuring the accurate administration of medications given to hospitalized patients. RFID technology also effectively monitors the location and movement of patients, hospital staff, and resources in the triage area in hospital emergency departments, and facilitates data collection and analysis.

At U.S. hospitals and medical centers—such as Massachusetts General Hospital, New York-Presbyterian Hospital, Georgetown University Hospital, Beth Israel Deaconess Medical Center, and Stanford University Medical Center—RFID solutions track patients and their physical medical records and monitor expiration dates of time-sensitive products including medications. In Korea, Won Ju Christian Hospital employs password-protected PDAs (personal digital assistants) as interrogators that work in concert with RFID tags embedded in wrist bracelets for mothers, and ankle bracelets for their infants, for securing newborns and eliminating mother and baby mix-ups at discharge (Dalton, Kim, & Lim, 2005).

To contain the outbreak of SARS (severe acute respiratory syndrome), RFID technology was used by health officials in Singapore to monitor the movements of every person entering and leaving Singapore General Hospital (Gopalakrishna, Choo, Leo, Tay, Lim, Khan, et al., 2004). Staff and visitors were required to wear credit card-sized RFD tags that transmitted their locations to sensors placed in hospital ceilings. As a consequence of the SARS outbreak, Singapore General Hospital also implemented an online physiotherapy program for enabling physical therapists to remotely monitor their patients, and teach them new exercises at their homes.

RFID tags on birds are used to contain outbreaks of avian flu. RFID tags on herds of cattle stem the spread of mad cow disease. The United States Department of Agriculture (USDA) supports a national registry that contains RFID data on infected livestock, such ad deer and bison. This information serves as a foundation for establishing quarantines to prevent widespread epidemics. Developed by the USDA, the National Animal Identification System (NAIS) also enables animal health officials to respond to emergencies and reduce disease spread.

PRIVACY AND SECURITY CONSIDERATIONS

RFID technology plays a vital role in monitoring the health and safety of patients in hospitals and medical centers. Nonetheless, the ability to obtain real-time detailed information as a consequence of RFID deployment also raises concerns about security (Nath, Reynolds, & Want, 2006). Possible abuse of RFID tracking capabilities also raises questions about potential violations of personal privacy (Ohkubo, Suzuki, & Kinoshita, 2005).

Generally, patient-related information collected by RFID systems in the healthcare space is extremely sensitive, and contains personal information that is protected by the Health Insurance Portability and Accountability ACT (HIPAA), and requires the enforcement of strict privacy controls (Karygiannis et al., 2006). Data obtained from RFID tags embedded in medical implantations and patients' wristbands for one purpose may be covertly used for monitoring individuals without their knowledge or consent. Data obtained from RFID tags embedded in consumer items such as shoes and clothing can potentially be used by employers to monitor surreptitiously the work of employees and terrorists to target attacks against specific political and ethnic groups. As a consequence of RFID system abuse and its potentially adverse impact on data confidentiality and privacy, groups such as CASPIAN (Consumers Against Supermarket Privacy Invasion and Numbers) have launched protest campaigns against manufacturers and retailers worldwide.

To counter these concerns, government entities such as Japan's Ministry of Economy, Trade, and Privacy, and consumer advocate organizations—including the Electronics Frontier Foundation and Electronic Privacy Information Center—have issued RFID privacy and security guidelines to safeguard the integrity of data collection. In California, legislation to impose limits on the use of RFID technology statewide has been proposed. According to the RFID Bill of Rights drafted by consumer advocates, consumers should be informed if products they purchase contain RFID tags, and be able to remove, deactivate, and/or destroy these tags and the data they contain. (Garfinkel, 2004). RFID privacy protection is also supported by encryption, key-exchange protocols and security specifications, such as the ITU X.509 standard.

Methods for disabling RFID tags to prevent responses to queries from interrogators are in development. A promising approach involves the use of clipped RFID tags that can be used by consumers to separate the RFID microchip from the RFID antenna, thereby deactivating RFID operations (Karjoth & Moskowitz, 2005). According to Juels, Rivest, and Szydlo (2003), utilization of selective blocking tags provides another option for addressing privacy concerns associated with the widespread use of RFID technology in consumer products. Selective blocking tags simulate ordinary RFID tags, but prevent RFID interrogators from reading RFID patient and consumer data. To address public opposition to RFID utilization, EPCglobal has also passed a standard that requires RFID tags to be equipped with a kill command, so that consumers can disable RFID functions. This specification is incorporated into the EPCglobal standard titled Class 1 Generation 2 UHF-Air Interface Protocol (Gen2). Additionally, the kill command also prevents utilization of the RFID tag after a product with a killed tag is resold or recycled (Ohkubo, Suzuki, & Kinoshita, 2005).

RECENT TRENDS

Healthcare advances that contribute to longevity also have resulted in the increase of debilitating age-related diseases including dementia and Alzheimer's. A new approach for enabling the elderly to remain at home involves utilization of sensor-augmented wireless sensor networks (WSNs) (Ho et al., 2005). In WSNs, self-powered sensors embedded into active and semiactive RFID tags are called sensor RFIDs. WSNs operate in the HF and UHF spectral frequencies and support a range of applications in the healthcare space, including the intake of medications by the elderly at regular intervals. By querying RFID tags placed on each pill container, WSNs detect changes in the amount of medicine available, and alert seniors to take their medications via blinking lights and beeping sounds. WSNs consisting of RFID tags affixed to everyday objects such as clothing, shoes, chairs, and beds can also track the daily living activities of the elderly at home and in independent living facilities. With the worldwide population expected to reach 761 million in 2025, WSNs are expected to play an increasingly important role in supporting elder healthcare solutions.

At Shin-chon Severance Hospital in Seoul, Korea, RFID-augmented WSNs monitor the temperature of blood bank refrigeration to ensure that this blood is safe to use in transfusions. RFID tags at Shin-chon Severance Hospital also track the location of blood bags in transit, so that the intended patient receives the assigned transfusion, thereby reducing the mortality risk for transfused patients and safeguarding the transfusion process (Brooks, 2005).

CONCLUSION

In the healthcare space, RFID technology is distinguished by its effectiveness in enabling patient identification, data collection, asset tracking, and real-time monitoring of the elderly at home and in independent living facilities (Philipose, Fishkin, Perkowitz, Patterson, Fox, Kautz, et al., 2005). In hospital environments, RFID implementations contribute to patient safety by eliminating incorrectly labeled medications, vials holding blood samples and laboratory specimen, and enable innovative application-specific services for treating infectious diseases such as SARS. RFID technology presents new options for healthcare enhancement, as well as new and ongoing security and privacy challenges that are not yet fully addressed.

REFERENCES

Brooks, J.P. (2005). Reengineering transfusion and cellular therapy processes hospitalwide: Ensuring the safe utilization of blood products. *Transfusion, 45*(4), 159–171.

Dalton, J., Kim, I.-H., & Lim, B.-K. (2005). *RFID technology in neonatal care*. Intel Corporation. Retrieved February 7, 2008, from http://www.intel.com/it/pdf/rfid2.pdf

Garfinkel, S. (2004). RFID rights. *MIT Technology Review*. Retrieved February 7, 2008, from http://www.technologyreview.com/Infotech/14278/

Garfinkel, S., Juels, A., & Pappu, R. (2005). RFID privacy: An overview of problems and proposed solutions. *IEEE Security & Privacy, 3*(3), 34–43.

Gopalakrishna, G., Choo, P., Leo, Y., Tay, B., Lim, Y., Khan, A., et al. (2004). SARS transmission and hospital containment. *Emerging Infectious Disease, 10*(3), 395–400.

Ho, L., Moh, M., Walker, Z., Hamada, T., & Su, C.-F. (2005). Applications: A prototype on RFID and sensor networks for elder healthcare: Progress report. In *Proceeding of the 2005 ACM SIGCOMM Workshop on Experimental Approaches to Wireless Network Design and Analysis* (pp. 70–75).

Juels, A., Rivest, R., & Szydlo, M. (2003). Selective blocking of RFID tags for consumer privacy. In *Proceedings of the 10th ACM Conference on Computer and Communications Security* (pp. 103–111).

Karjoth, G., & Moskowitz, P. (2005). Disabling RFID tags with visible confirmation: Clipped tags are silenced. In *Proceedings of the 2005 Workshop on Privacy in the Electronic Society* (pp. 27–30).

Karthikeyan, S., & Nesterenko, M. (2005). RFID security without extensive cryptography. In *Proceedings of the Third Annual ACM Workshop on Security of Ad Hoc and Sensor networks* (pp. 63–67).

Karygiannis, T., Eydt, B., Barber, G., Bunn, L., & Phillips, T. (2006). *Guidance for securing radio frequency identification (RFID) systems. Recommendations of the National Institute of Standards and Technology*. Special publication (pp. 800–998) (Draft). U.S. Department of Commerce: National Institute of Standards and Technology.

Littman, M.K. (2002). *Building broadband networks*. Boca Raton, FL: CRC Press.

Myung, J., & Lee, W. (2006). Adaptive binary splitting: A RFID tag collision arbitration protocol for tag identification. *Mobile Networks and Applications*, *11*, 711–722.

Nath, B., Reynolds, F., & Want, R. (2006). RFID Technology and applications. *IEEE Pervasive Computing, 5*(1), 22–24.

Oak Ridge National Laboratory. (2007). Radio frequency identification (RFID) protected asset tracking system (PATS). *Engineering Science and Technology Division Internal Newsletter*. Retrieved February 7, 2008, from http://www.ornl.gov/sci/engineering_science_technology/technical_article_for_public_site/RFID/radio.shtml

Ohkubo, M., Suzuki, K., & Kinoshita, S. (2005). RFID privacy issues and technical challenges. *Communications of the ACM, 48*(9), 66–71.

Philipose, M., Fishkin, K., Perkowitz, M., Patterson, D., Fox, D., Kautz, H., et al. (2004). Inferring activities from interactions with objects. *IEEE Pervasive Computing, 3*(4), 50–57.

Philipose, M., Smith, J., Jiang, B., Mamishev, A., Roy, S., & Sundara-Rajan, K. (2005). Battery-free wireless identification and sensing. *IEEE Pervasive Computing, 4*(1), 37–45.

Staake, T., Thiesse, F., & Fleisch, E. (2005). Extending the EPC network – the potential of RFID in anti-counterfeiting. In *Proceeding of the 2005 ACM SIGCOMM Workshop on Experimental Approaches to Wireless Network Design and Analysis* (pp. 1607–1612).

Want, R. (2004). Que focus: RFID: The magic of RFID. *Que, 2*(7), 41–48.

Weinstein, R. (2005). RFID: A technical overview and its application to the enterprise. *IT Professional, 7*(3), 27–33.

KEY TERMS

Active RFID Tag: Features an onboard battery that serves as its own power source for performing operations.

Application-Specific Service: Service that satisfies the performance and functional requirements of either an application or class of applications.

Barcode: Consists of lines of different widths that identify an item. Works with a scanner.

Passive RFID Tag: Lacks an onboard power source. Makes use of incoming radio waves broadcast by an interrogator to power its response.

Radio Frequency Identification (RFID): Wireless identification and data capture technology that consists of a tag or transponder and an interrogator or reader to support applications ranging from airport baggage handling to supply chain management.

RFID System: Consists of the RFID tag, the RFID reader, and the communications between them.

Improving Consumer Health Literacy with Information Technology

Gondy Leroy
Claremont Graduate University, USA

INTRODUCTION

Before the Internet became popular as a device for distributing and sharing information, people turned to friends, books, and their doctors when they had medical questions. Today, many more options exist (Figure 1). Hundreds of Web sites provide health information and opportunities for interaction among patients, doctors, and caregivers. Estimates differ, but all surveys show that millions of people search online for health information. A Pew survey estimates that 80% of adult Internet users, about 93 million Americans, searched online for at least one of 16 major health topics (Fox & Fallows, 2003). Baker, Wagner, Signer, and Bundorf (2003) estimate that 20% of the U.S. population uses the Internet to find health information. A larger proportion (71%) of older people (50 to 64 years old) compared to 53% of younger people (18 to 29 years old) turn to the Internet for health information (Fox & Rainie, 2002). Although there is a digital divide, use of information technology is not simply decided by race or social class. Safran (2003) found that Medicaid families, who are believed not to use these new technologies, accessed their online Baby CareLink from the hospital, work, library, or other public access points. Gustafson et al. (2002) point out that poverty is the prime indicator for lack of technology use.

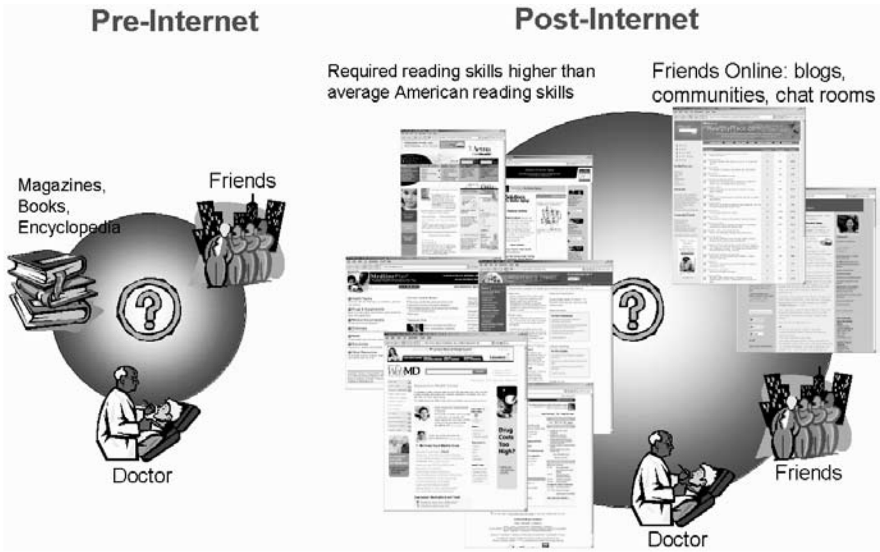
The problem we address in this chapter is consumers' lack of understanding of the available information. This is extremely important since for at least a third of these consumers, the information affects decisions about their health, health care, and visits to a health care provider (Baker et al., 2003). Warner and Procaccino (2004) found a much higher percentage in his interviews with women; more than 80% responded that the information they found online affected their decisions about treatments.

BACKGROUND

Thousands of Web sites provide information and additional opportunities to share information in an interactive format. The information can be targeted at the general public or a specific subgroup, and there are several advantages to this trend. Foremost, consumers will be more informed. This is a benefit because it empowers them to ask more informed questions when seeing their caregivers and lessens their fears of the unknown (Fox & Fallows, 2003). Often, physicians want to refer their patients to Web sites and printed patient educational material for additional information (Brawn, 2005). The online information is especially beneficial for consumers who need more detailed information than their health care provider can give in a limited amount of time. For example, Rosmovits and Ziebland (2004) conducted in-depth interviews with cancer patients and found that they have complex information needs that were not met by their health care providers. They felt they received incomplete and sometimes contradictory information from their caregivers. Consumers also interact with each other online to provide information and support. There are many support groups where members share advice or provide support in difficult times (e.g., multiple sclerosis patients supporting each other during painful self-injections) (Johnson & Ambrose, 2006).

Unfortunately, there are also disadvantages associated with health information as it is currently provided online. The disadvantages can be classified into two groups related to incorrect information and incorrect understanding of information. Since the Internet is not regulated, there is no guarantee that the information provided is correct and trustworthy. The general public should be educated in the usage of this information. Murray, et al. (2003) questioned physicians and found that 75% of the respondents felt that health information on the Internet was a good thing. However, the quality

Figure 1. Information sources for consumers

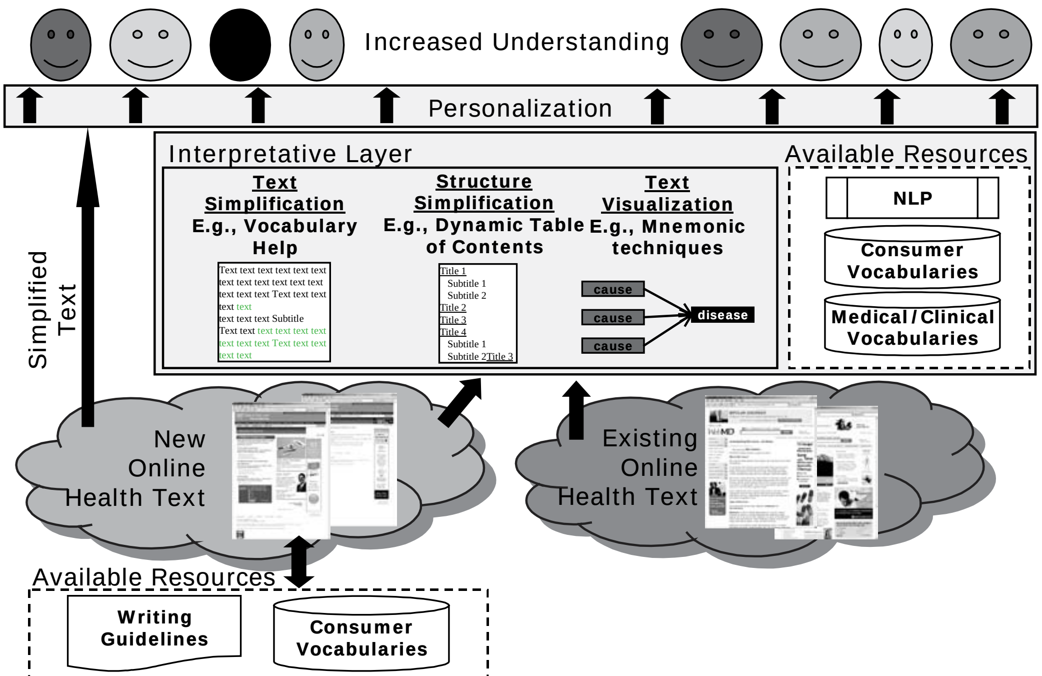


of online information affected the health care outcome and the patient-physician relationship. Accurate and relevant information had a beneficial effect on both. In addition, the outcome and relationship were also influenced by the physician's perceived threat to his or her authority, especially when the patient wanted something inappropriate. The second group of problems centers on health consumers' lack of understanding of this information (Berland et al., 2001; D'Alessandro, Kingsley, & Johnson-West, 2001; Root & Stableford, 1999) and has consequences for health care at large. The Ad Hoc Committee on Health Literacy for the Council on Scientific Affairs (1999) found that misunderstandings in health information increase the risk of making unwise health decisions leading to poorer health and higher health care costs. For example, Garbers and Chiasson (2004) showed that Latinas with low health literacy were significantly less likely to have preventive cervical cancer screening. Kalichman, Benotsch, Suarez, Catz, Miller, and Rompa (2000) found that HIV/AIDS patients with low literacy levels were more likely to (incorrectly) believe that antiviral drugs would help prevent transmitting HIV during unprotected sex.

Figure 2 shows how technology can help improve understanding of health information. Current research is still in the early stages and has not advanced much beyond measuring reading levels and describing these in numerous studies of consent forms and patient leaflets for a variety of afflictions. Existing interventions focus on tailoring information to specific groups (tailored information) or individual people (targeted intervention) (McCray, 2005). In general, newly written text should take writing guidelines into account, and simple versions of the material should be available where possible. There are several guidelines that can be consulted: MedlinePlus (http://www.nlm.nih.gov/medlineplus/etr.html) provides guidelines for writing easy-to-read versions of documents; the National Institutes of Health (NIH) provide the Plain Language Initiative (http://execsec.od.nih.gov/plainlang/index.html); the state of California provides the California Health Literacy initiative (http://cahealthliteracy.org/); and the Health & Literacy Special Collection (http://lincs.worlded.org/) also provides advice. Regrettably, it is impossible to provide simplified versions of all information because the content itself may be too complex. It is also infeasible to rewrite all existing text even if one were to limit it to English. Automated tools need to be developed to help consumers understand text.

Three approaches can be followed and combined to help increase understanding of the information. First, the language and grammar used in the text can be simplified (text simplification). Second, the structure of the text can be visualized and simplified so the document is easier to follow (structure simplification). Finally, important information can be simplified, visualized, and emphasized (text visualization). These techniques

Figure 2. Making health information understandable for consumers



should be partially or completely automated and optimized for various consumer groups.

In the following, an overview is provided of current research that fits into this framework. The focus is on English language Web sites. Then, key consumer groups who would benefit most are discussed. Finally, future trends are described.

INTERPRETATIVE LAYER

Text Simplification

Several formulas are commonly used to measure readability (Berland et al., 2001; D'Alessandro et al., 2001; Root & Stableford, 1999), and all studies using them show that a significant portion of information is too difficult for average adults to read. Most evaluation studies use the Flesch readability scores or the Flesch-Kincaid grade levels to evaluate text. These formulas use syntax, word counts, and word length to assign readability levels and are easily available with Microsoft Word. An additional popular measure is the SMOG measure (McLaughlin, 1969), which is based on syllable count. Freda (2005), however, found the SMOG measure assigned reading levels 2 or 3 grades higher than the Flesch-Kincaid grade levels. Most English sites require at least a 10th grade (Flesch) reading level, and more than half present information at college level. This is perhaps a partial explanation of the fact that Internet usage for health information is strongly associated with higher education (Baker et al., 2003; Fox & Fallows, 2003).

Additional metrics exist to approach the problem from the consumer side and focus on a person's health literacy level instead of the difficulty of the text. For example, the Test of Functional Health Literacy in Adults (TOFHLA), its shortened version STOFHLA (Parker, Baker, Williams & Nurss, 1995) and the Cloze procedure (Taylor, 1953) are often used. The Rapid Estimate of Adult Literacy in Medicine (REALM) (Davis et al., 1993) is one of the most popular measures (Pignone, DeWalt, Sheridan, Berkman, & Lohr, 2005). These metrics provide a common method to evaluate consumer understanding of common medical and health terminology and have led to surprising findings. For example, Zun, Sadoun, and Downey (2006) found that nurses and doctors overestimate written English competency in their Hispanic patients when compared to test results with the REALM and STOFHLA. Although these measures are commonly used, few studies directly compare performance of individuals (health literacy) with different text reading levels (readability measures). A recent exception is the work by Trifiletti, Shields, McDonald, Walker, and Gielen (in press) who evaluated text of different grade levels and performed the Cloze test on these texts. The Cloze test removes the n^{th} word from a text and then requires people to fill in the blanks. They found that more people performed at an acceptable level with lower grade level texts. Similarly, Pignone, et al. (2005) provide a review of studies focusing on text simplification and showed improved understanding in the low literacy group when texts were simplified.

There are two components that can be focused on to automatically simplify texts and lessen the aforementioned difficulties. The first is grammar; the second is vocabulary. Grammar and sentence structure simplification would lower the readability grade levels. For example, using active instead of passive sentences or right branching instead of embedded or left branching sentences increases readability. However, simply lowering the required reading level will be insufficient. Leroy, Eryilmaz, and Laroya (2006) found that difficult documents measured with the Flesch readability formulas not only use more complex sentence structures but also use more complex vocabulary and often discuss more difficult topics. These results correspond with Boulos' (2005) conclusion that some documents will remain difficult, and other means of user support will be needed.

Structure Simplification

Most research has concentrated on visualizing large collections of entire documents or visualizing information extracted from all those documents. Little research focuses on facilitating content access for a single document and even less on medical or health text. However, previews of single documents have been found to expedite review of documents (Greene, Marchionini, Plaisant & Shneiderman, 2000; Marchionini, Plaisant & Komlodi, 1998). A preview extracted from the original document acts as a surrogate. It is effective when it communicates sufficient information to the user about individual information items. For example, most search engines provide an excerpt of text called a "snippet." Some simply show the first few lines of text; others display the text surrounding keywords, use heuristics to select sections of documents (Amitay & Paris, 2000), or provide a document summary.

The previews can be textual, graphical, or a combination of both. Woodruff, Rosenholtz, Morrison, Faulring, and Pirolli (2002) compared textual and graphical (thumbnail) summaries and found the best results with a combination of both types. Manber (1997) combined color with the original text and proposes highlighting as a preview for documents. By keeping a personal list of keywords and highlighting these in a document, users can rapidly determine their interest in it. Hornbæk & Frøkjær (2001, 2003, 2004) studied linear, fisheye, and overview+detail interfaces to facilitate reading of electronic documents. They compared their subjects' essay, question-answer task performance, and user satisfaction. They manipulated how much text was visible to users at any time but did not extract any text or information from the text. Even so, the overview+detail interface helped their subjects understand main ideas better, and the fisheye interface helped them answer questions faster. The linear text interface was worse than the other two in most aspects. This approach is rarely evaluated for health information, with a few exceptions. Ogozalek (1994) used text and multimedia interfaces to provide prescription drug information to the elderly and found that the subjects answered questions better with the multimedia interface. There was no effect on retention of the information. Miller, Leroy, and Wood (2006) are working on dynamically generating tables of contents for WebMD documents using UMLS semantic types as entry points.

Text Visualization

More advanced approaches extract and visualize pertinent information from individual documents. This approach is being tested for biomedical text and biomedical researchers. For example, in Genescene (Leroy & Chen, 2005), an interactive graph visualization of biomedical relationships is shown. Researchers can browse the graphs and drill down to the underlying text. A simpler approach is used by Beier and Tesche (2001), who provide an overview of a metasearch based on the origin of the results, such as Internet sites or journals. However, no such tools are currently tested for or made available to health information consumers.

HEALTH INFORMATION CONSUMER GROUPS

Improved access, understanding, and retention of health information would benefit the general population. However, three groups in particular deserve special attention: non-native English speakers, the elderly, and patients. These three groups have special health information needs, which we will discuss in more detail, and are a growing group online. For example, Gustafson et al. (2002) focused on underserved African Americans, the elderly, and HIV patients, and found that these three groups tend to use mostly information and analysis services and not so much communication technology such as the discussion boards. As such, increasing their understanding of online health texts should be a key focus.

Many online readers who do not speak English as their native language still read information in English. To fully appreciate the problem, think about a second or third language you speak and how difficult it would be to understand the health information in that language. This is problematic and affects health care outcomes. The effect will be stronger when information is only available in English. In addition to the text, several other factors, such as cultural differences, structure of pamphlets, difficult images, and lack of definitions, affect health literacy (Hunter, 2005). Few realistic and feasible approaches are proposed in the literature. Popular advice is to provide translations of all text. However, Becker (2004) found that only 10% of state sites in the United States provide Spanish translations, and many buttons and links are never translated. Moreover, Parker and Kreps (2005) illustrated that translating all texts is not a scalable solution since health care organizations may need translations in 40 or more languages. Translations will not be available for everyone, and so this multilingual consumer group will benefit most from text simplification.

A second growing online group is the elderly. There have been large survey-based studies that looked at the relationship between Internet usage and demographic variables, such as race, gender, and age. Ito, O'Day, Adler, Linde, and Mynatt (2001) studied SeniorNet, an online community with more than 20,000 members and 4,000 volunteers. They found that seniors do not see themselves as different or technologically challenged. Even so, with increasing age, people encounter problems that may interfere with optimal Internet usage. Vision deteriorates, requiring bigger font sizes and more contrasting colors. Mouse and keyboard skills are often lower due to physical problems such as arthritis, tremors, or lack of experience with computers in general. Learning becomes slower and more difficult with longer training times and more attention problems (Hanson, 2001; Nielsen, 2000). This group may benefit especially from improved text structure since they perform as well as younger adults in recalling stories when events are in canonical order (Wingfield & Stine-Morrow, 2000). Although elder users do not see themselves as less experienced, Chadwick-Dias, Tedesco, and Tullis (2004) demonstrate that self-reported Web experience is not the same as actual Web expertise. Older users received lower scores on an expertise quiz, even when controlled for self-reported Web experience. The quiz consisted of a list of images and their possible meanings (e.g., a back button).

Patients are a third growing online group. This group is very diverse, and different types of patients have different needs. Most representative of this group are individuals who need information on a recently diagnosed disease. Depending on the seriousness of the illness, these patients are more or less stressed, tired, and fearful. All these factors may affect how well they grasp health information. For example, Van Servellen, Brown, Lombardi, and Herrera (2003) found that increased stress was significantly correlated with poorer recognition of HIV terms in their group of low-income Latinos. Most patients in this group will not understand all the vocabulary and information presented online. In contrast, there is an additional group of expert patients who have different needs. Although there are different opinions on what an expert patient is (Shaw & Baker, 2006), in the best scenario, these are patients with chronic diseases who are involved in the management of their disease and who collaborate with health care providers for the best outcome (Badcott, 2005). This expert patient group will have more background knowledge, use a much more advanced vocabulary, and understand more complicated documents.

FUTURE TRENDS

In the future, two broad trends can be expected. The first will be studies that focus mainly on user groups. There will be an increasingly sophisticated knowledge base that can distinguish between various user groups and their ability to learn and remember information in interaction with various types of technology. Both cognitive science and educational methodologies will play a significant role. Although this chapter focused on adults, children will become a much more important consumer group.

The second trend will be complementary and will comprise various types of technology and media and their interaction with various consumer groups. Computational linguistics and computer and information science will drive these trends. For example, text simplification will benefit consumers directly, but may also lead to better machine translation. Visualization can be used to visualize the structure of text, but also for treatments and interventions. Text visualization may be adapted and act as summaries or previews. PDA or Smartphone users with smaller screens would particularly benefit from previews. Such previews may also become relevant for medical professionals in the field when dealing with emergencies that require them to find new information. Voice recognition combined with wireless search engine technology to retrieve information and show this on small displays (e.g., watches, glasses) will also become available. More advanced visualization such as automated translation from text to graphical novels or movies and animation would be especially beneficial for low literacy groups and children.

CONCLUSION

Understanding health information is an important aspect of our health care. It leads to more informed and more comfortable patients and better adherence to therapies. Many studies show that the general public does not understand the information from which they would benefit. Current research to facilitate this understanding is in the early stages. Most research has focused on demonstrating the problem. Some are now evaluating how to simplify the information, but approaches are manual. To tackle the problem in an efficient manner, automated approaches are needed that can scale up. This research is in its infancy, but the benefits will be enormous as it matures. More and better-informed health consumers from all walks of life will benefit. And research which focuses on related problems such as summarization, visualization, or even machine translation may also be affected.

ACKNOWLEDGMENT

This work is funded by a grant from the National Library of Medicine, R21-LM008860-01, Visualization of Consumer Health Information, http://isl.cgu.edu/ConsumerHealth.htm.

REFERENCES

Ad Hoc Committee on Health Literacy for the Council on Scientific Affairs, A.M.A. (1999). Health literacy: Report of the Council on Scientific Affairs. *JAMA, 281,* 552–557.

Amitay, E., & Paris, C. (2000). *Automatically summarising Web sites: Is there a way around it?* Proceedings of the Ninth International Conference on Information and Knowledge Management, McLean, Virginia.

Badcott, D. (2005). The expert patient: Valid recognition or false hope? *Medicine, Health Care Philosophy, 8*, 173–178.

Baker, L., Wagner, T.H., Signer, S., & Bundorf, M.K. (2003). Use of the Internet and e-mail for health care information: Results from a national survey. *Journal of the American Medical Association, 289*(18), 2400–2406.

Becker, S.A. (2004). A study of Web usability for older adults seeking online health resources. *ACM Transactions on Computer-Human Interaction, 11*(4), 387–406.

Beier, J., & Tesche, T. (2001). Navigation and interaction in medical knowledge spaces using topic maps. *International Congress Series, 1230*, 384–388.

Berland, G.K., et al. (2001). Health information on the Internet: Accessibility, quality, and readability in English and Spanish. *Journal of the American Medical Association, 285*, 2612–2621.

Boulos, M.N.K. (2005). British Internet-derived patient information on diabetes mellitus: Is it readable? *Diabetes Technology & Therapeutics, 7*(3), 528–535.

Brawn, T.S. (2005). Consumer health libraries: What do patrons really want? *Journal of the Medical Library Association, 93*(4), 495–496.

Chadwick-Dias, A., Tedesco, D., & Tullis, T. (2004). *Older adults and Web usability; Is Web experience the same as Web expertise?* Proceedings of the CHI-2004, Vienna, Austria.

D'Alessandro, D., Kingsley, P., & Johnson-West, J. (2001). The readability of pediatric patient education materials on the World Wide Web. *Archives of Pediatrics and Adolescent Medicine, 155*, 807–812.

Davis, T., et al. (1993). Rapid estimate of adult literacy in medicine: A shortened screening instrument. *Family Medicine, 25*(6), 391–395.

Fox, S., & Fallows, D. (2003). *Internet health resources—Health searches and email have become more commonplace, but there is room for improvement in searches and overall Internet access.* Washington, DC: Pew Internet & American Life Project.

Fox, S., & Rainie, L. (2002). *Vital decisions: How Internet users decide what information to trust when they or their loved ones are sick.* Washington, DC: Pew Internet & American Life Project.

Freda, M.C. (2005). The readability of American Academy of Pediatrics patient education brochures. *Journal of Pediatric Health Care, 19*(3), 151–156.

Garbers, S., & Chiasson, M.A. (2004). Inadequate functional health literacy in Spanish as a barrier to cervical cancer screening among immigrant Latinas in New York City. *Preventing Chronic Disease, 1*(4), 1–10.

Greene, S., Marchionini, G., Plaisant, C., & Shneiderman, B. (2000). Previews and overviews in digital libraries: Designing surrogates to support visual information seeking. *Journal of the American Society for Information Science, 51*(4), 380–393.

Gustafson, D.H., et al. (2002). CHESS: 10 years of research and development in consumer health informatics for broad populations, including the underserved. *International Journal of Medical Informatics, 65*, 169–177.

Hanson, V.L. (2001). *Web access for elderly citizens.* Proceedings of the 2001 EC/NSF Workshop on Universal Accessibility of Ubiquitous Computing: Providing for the Elderly.

Hornbæk, K., & Frøkjær, E. (2001). Reading of electronic documents: The usability of Linear, Fisheye, and Overview+Detail interfaces. In *Proceedings of the SIGCHI '01*, Seattle, Washington.

Hornbæk, K., & Frøkjær, E. (2003). Reading patterns and usability in visualizations of electronic documents. *ACM Transactions on Computer-Human Interaction, 10*, 119–149.

Hornbæk, K., & Frøkjær, E. (2004). Reading patterns and usability in visualizations of electronic documents. *Interactions, 11*(1), 11–12.

Hunter, J.L. (2005). Cervical cancer educational pamphlets: Do they miss the mark for Mexican immigrant women's needs? *Cancer Control, 12* (Cancer, Culture and Literature Supplement), 42–50.

Ito, M., O'Day, V.L., Adler, A., Linde, C., & Mynatt, E.D. (2001). Making a place for seniors on the net: SeniorNet, senior identity, and the digital divide. *Computers and Society*, 15–11.

Johnson, G.J., & Ambrose, P.J. (2006). Neo-tribes: The power and potential of online communities in health care. *Communications of the ACM, 49*(1), 107–113.

Kalichman, S.C., Benotsch, E., Suarez, T., Catz, S., Miller, H., & Rompa, D. (2000). Health literacy and health related knowledge among persons living with HIV/AIDS. *American Journal of Preventive Medicine, 18*(4), 325–331.

Leroy, G., & Chen, H. (2005). Genescene: An ontology-enhanced integration of linguistic and co-occurrence based relations in biomedical texts. *Journal of the American Society for Information Science and Technology (Special Issue), 56*(5), 457–468.

Leroy, G., Eryilmaz, E., & Laroya, B.T. (2006). *Health information text characteristics.* Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium, Washington, DC.

Manber, U. (1997). *The use of customized emphasis in text visualization.* Proceedings of the IEEE Conference on Information Visualization.

Marchionini, G., Plaisant, C., & Komlodi, A. (1998). Interfaces and tools for the Library of Congress National Digital Library Program. *Information Processing and Management, 34*(5), 535–555.

McCray, A.T. (2005). Promoting health literacy. *Journal of the American Medical Informatics Association, 12*(2), 152–163.

McLaughlin, G.H. (1969). SMOG grading: A new readability formula. *Journal of Reading, 12*, 636–646.

Miller, T., Leroy, G., & Wood, E. (2006). *Dynamic generation of a table of contents with consumer-friendly labels.* Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium, Washington, DC.

Murray, E., et al. (2003). The impact of health information on the Internet health care and the physician-patient relationship: National U.S. survey 1.050 physicians. *Journal of Medical Internet Research, 5*(3), 17.

Nielsen, J. (2000). *Designing Web usability*. New Riders Publishing.

Ogozalek, V.Z. (1994). *A comparison of the use of text and multimedia interfaces to provide information to the elderly.* Proceedings of the CHI, Boston, Massachusetts.

Parker, R., Baker, D., Williams, M., & Nurss, J.R. (1995). The test of functional health literacy in adults: A new instrument for measuring patients' literacy skills. *Journal of General Internal Medicine, 10*(10), 537–541.

Parker, R., & Kreps, G.L. (2005). Library outreach: Overcoming health literacy challenges. *Journal of the Medical Library Association, 93*(4), S81–S85.

Pignone, M., DeWalt, D.A., Sheridan, S., Berkman, N., & Lohr, J.N. (2005). Interventions to improve health outcomes for patients with low literacy: A systematic review. *Journal of General Internal Medicine, 20*(2), 185–193.

Root, J., & Stableford, S. (1999). Easy-to-read consumer communications: A missing link in Medicaid managed care. *Journal of Health Politics, Policy, and Law, 24*, 1–26.

Rosmovits, L., & Ziebland, S. (2004). What do patients with prostate or breast cancer want from an Internet site? A qualitative study of information needs. *Patient Education and Counseling, 53*(1), 57–64.

Safran, C. (2003). The collaborative edge: Patient empowerment for vulnerable populations. *International Journal of Medical Informatics, 69*, 185–190.

Shaw, J., & Baker, M. (2006). "Expert patient"—Dream or nightmare? *British Medical Journal, 328*(27).

Taylor, W.L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly, 30*, 415–433.

Trifiletti, L.B., Shields, W.C., McDonald, E.M., Walker, A.R., & Gielen, A.C. [in press]. Development of injury prevention materials for people with low literacy skills. *Patient Education and Counseling*.

Van Servellen, G., Brown, J.S., Lombardi, E., & Herrera, G. (2003). Health literacy in low-income Latino men and women receiving antiretroviral therapy in community-based treatment centers. *Aids Patient Care and STDs, 17*(6), 283–298.

Warner, D., & Procaccino, J.D. (2004). Toward wellness: Women seeking health information. *Journal of the American Society for Information Science and Technology, 55*(88), 709–730.

Wingfield, A., & Stine-Morrow, E.A.L. (2000). Language and speech. In T.A. Salthouse (Ed.), *The handbook of aging and cognition* (pp. 359–416). Mahwah, NJ: Lawrence Erlbaum Associates.

Woodruff, A., Rosenholtz, R., Morrison, J.B., Faulring, A., & Pirolli, P. (2002). A comparison of the use of text summaries, plain thumbnails, and enhanced thumbnails for Web search tasks. *Journal of the American Society for Information Science and Technology, 53*(2), 172–185.

Zun, L.S., Sadoun, T., & Downey, L. (2006). English-language competency of self-declared English-speaking Hispanic patients using written tests of health literacy. *Journal of the National Medical Association, 98*(6), 912–917.

KEY TERMS

Cloze Procedure: A procedure to measure readability of text by deleting the n^{th} word and asking readers to fill in the blanks. This procedure was originally developed by Taylor (1953).

Consumer Health Informatics: Research that focuses on health information, the consumers who read the information, and the interaction between the two.

Flesch Reading Ease: A readability measure that provides an estimate of the readability of a text and the required grade level to understand the text. It is based on syntactic and word level considerations.

Health Literacy: The ability to understand health and medical information and act upon that information correctly.

Readability: The difficulty level of a text, usually measured by formulas such as the Flesch Reading Ease.

REALM: Rapid Estimate of Adult Literacy in Medicine.

STOHFLA: Short Test of Functional Health Literacy in Adults (Davis et al., 1993).

TOFHLA: Test of Functional Health Literacy in Adults (Parker et al., 1995).

Incremental Neural Network Training for Medical Diagnosis

Sheng-Uei Guan
Xian Jiatong-Liverpool University, China

Ji Hua Ang
National University of Singapore, Singapore

Kay Chen Tan
National University of Singapore, Singapore

Abdullah Al Mamun
National University of Singapore, Singapore

INTRODUCTION

Artificial intelligence has attracted many researchers in recent years for its usefulness in the medical field. With the increase in research interest, artificial intelligence systems have found widespread usage in the medical domain (Remzi & Djavan, 2004; Tsakonas, Dounias & Jantzen, 2004). These applications arise as medical datasets usually contain a high volume of data, and it is very costly for human manual analysis and handling. The intelligent systems posses the ability to process datasets, extract useful information from them, and interpret the data at a much lower cost (Raymer, Doom, Kuhn & Punch, 2003) as compared to the manual handling. Intelligent systems like evolutionary computing, fuzzy logic, and neural networks are just some that many researchers have recently looked into.

Neural network has been used in the medical field to a great extent (Hayashi & Setiono, 2002) and more noticeably for classification problems (Abbass, 2002), since more frequently the outcome of the medical diagnosis is to categorize the patient into the different classes. The conventional method of neural network training for medical diagnosis is to present all the input attributes together as a batch at the same time to the neural network. However, some input attributes are more important and contain a higher consideration factor than others, and the criteria that each attribute uses for classification is also different. If all inputs are presented as a batch at the same time, trouble arises easily if there is interference among them. Thus, if all the input attributes are trained together in a batch, each attribute will interfere in the decision-making of others.

The input feature space and architecture of the output layer has proved to be a factor affecting the performance of neural network (Guan & Li, 2002a; Guan & Li, 2002b; Guan & Li, 2004; Guan, Li, & Qi, 2004). Input data space clustering techniques were applied in neural networks studies (Plikynas, 2004; Sun & Peterson, 1999). Sun and Peterson (1999) partitioned the input space into different regions and applied differential weighting for different regions so they have different agents that specialize in local regions. However, to date, little work has been reported to investigate the interference of the input attributes and, inferring from the information, to devise a carefully designed feature-partitioning algorithm.

This chapter proposes a novel method of incremental interference-free neural network training (IIFNNT) for medical datasets, which takes into consideration the interference each attribute has on the others. A specially designed network is used to determine if two attributes interfere with each other, after which the attributes are partitioned using some partitioning algorithms. These algorithms make sure that attributes beneficial to each other are trained in the same batch, thus sharing the same subnetwork while interfering attributes are separated to reduce interference. There are several incremental neural networks available in literature (Guan & Li, 2001; Su, Guan & Yeo, 2001). The architecture of IIFNNT employed some incremental algorithm: the ILIA1 and ILIA2 (incremental learning with respect to new incoming attributes) (Guan & Li, 2001).

This chapter is divided into five sections. The next section describes the incremental neural networks used in the IIFNNT. In Section 3, the method to evaluate interference between two attributes and the interference-free attribute partitioning algorithm is presented. The results of experiments on a diabetes medical dataset are given and analyzed in Section 4, together with a comparison to other related work in literature. The conclusion of our research, together with suggestions for future work, is given in Section 5.

Incremental Neural Networks Used

As mentioned earlier, the architecture of IIFNNT is based on the incremental algorithm ILIA (Guan & Li, 2001). When additional input attributes are considered, the ILIA expands its input space to accommodate these new attributes. The architecture of the ILIA algorithms from Guan and Li (2001) is described here.

ILIA1

The existing available attributes are first presented to the network, as shown in Figure 1. Network consists of direct connections from all the input units to all the output units. All input units are connected to all the hidden units. All the hidden units are also connected to all output units. The ILIA used the Constructive Backpropagation Learning algorithm (Lehtokangas, 1999) to train the weights and to determine the number of hidden neurons needed for its network. After the weights of the network have been trained, they are fixed and no longer changed.

When there are new attributes to be considered, increase the input dimension to accommodate the new attributes and a new subnetwork (in terms of new direct connections and new hidden units) is added to the existing network. The newly formed neural network architecture is shown in Figure 2. The process of adjusting the weights and installing new hidden units is done only for the new subnetwork.

ILIA2

An extension to ILIA1 is also given in Guan and Li (2001). In ILIA2 (Figure 3), new output units are added to the neural network obtained in ILIA1. The number of new output units is equal to the number of output units in ILIA1. Thus, with new output units being added to the neural network obtained in ILIA1, the output units in ILIA1 are being collapsed and have effectively become a new hidden layer.

INTERFERENCE-FREE NEURAL NETWORK

There are several steps needed to form the interference-free neural network. The steps will be broken down and the major ones presented in the following subsections. An example using the diabetes dataset will also be used along the way to illustrate the method.

Figure 1. Initial network

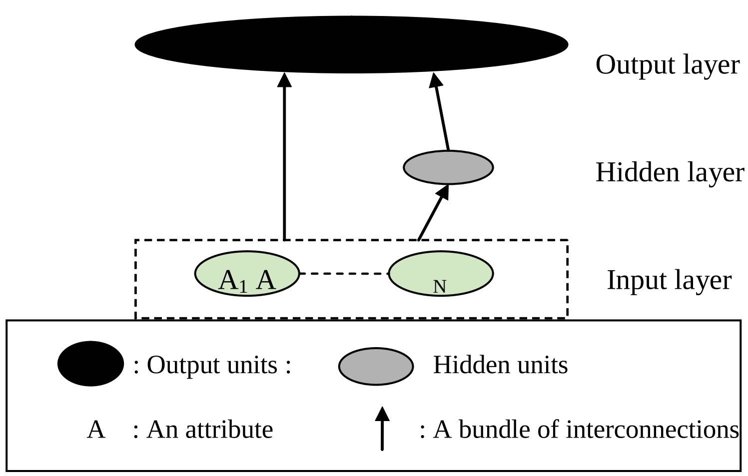


Figure 2. Architecture of ILIA1

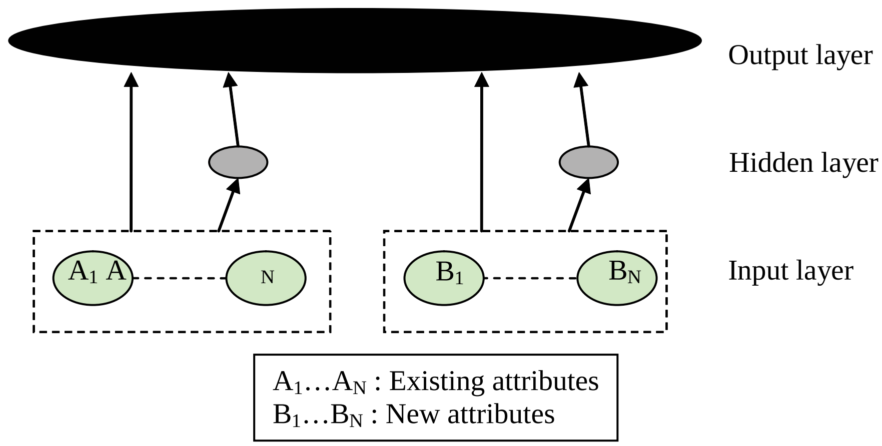
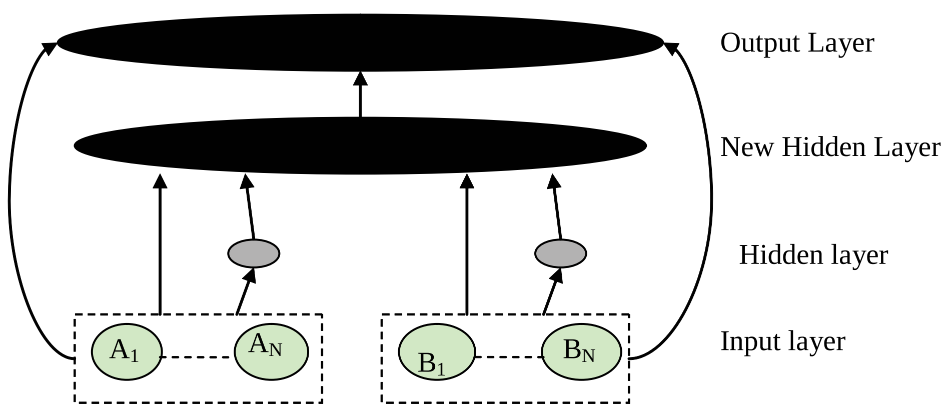


Figure 3. Architecture of ILIA2



Formulating the Interference Table

The interference-free partitioning method is inferred from the Interference Table. The Interference Table consists of individual attribute discrimination ability (Guan & Li, 2004) and the discrimination ability of any two attributes when trained together.

Evaluating Individual Attribute Discrimination Ability

Guan & Li (2004) state an attribute's individual discrimination ability as the ability of each individual input attribute to correctly classify a given problem. The assessment of such ability is by having just an attribute (the attribute to be tested for its discrimination ability) in the input space of the neural network for training. This is done in the absence of other attributes. The network used here is similar to the network in Figure 1; the difference is that there is only one input attribute in this case, whereas the network in Figure 1 can take any number of attributes.

When training this network, all the patterns in the training dataset are used to train the neural network. There is only one attribute available in the training dataset, which is the attribute to be tested. The ILIA (Guan & Li, 2001) referred to Prechelt's (1994) work and used the method of *early stopping* for the Constructive

Backpropagation Learning Algorithm (Lehtokangas, 1999). The validation set is used here for evaluating the quality of the network during training and to prevent overfitting.

The error produced on the validation set at the end of the training will be a measurement of the individual attribute discrimination ability. The error is tabulated in Table 1.

Evaluating Discrimination Ability of Two Attributes When Trained Together

The neural network architecture shown in Figure 4 is used to evaluate the discrimination ability when any two attributes are trained together. At any one time, there are only two input attributes in the input layer. Once again, this network is similar to the network, as shown in Figure 1. The difference is that there can only be two input attributes (the two attributes to be tested) in the input layer.

There will be two attributes available in the training and validation set, and all the patterns in the training and validation set will be used for the training process. Output of the network will be decided by these two attributes trained in batch (both input attributes have connections to the same hidden units), and consequently, the percentage of error produced on the validation set is noted down and tabulated in the Interference Table.

The Interference Table

Formulation of this table is able to illustrate whether there exists interference between any two attributes. Table 1 gives an example of how an Interference Table looks. By comparing the error produced from training the two attributes together, I_{ij} with error produced by each attribute when trained alone, IDA_i or IDA_j it will be able to show if interference exists among these attributes. Two attributes are said to be interfering with each other if the error derived from training both the attributes in batch is higher than the error resulted from training either attribute alone (i.e., $I_{ij} > IDA_i$ or $I_{ij} > IDA_j$). Two attributes are not interfering if the error derived from training both the attributes in batch is lower or equal than the error resulted from training either attribute alone (i.e., $I_{ij} < IDA_i$ and $I_{ij} < IDA_j$). If error does not increase or decrease when two attributes are trained in batch compared to training them alone (i.e., $I_{ij} = IDA_i$ and $I_{ij} = IDA_j$), these two attributes most likely do not have an effect on one another.

For clarity, the interference table of the diabetes set is given in Table 2. For example, the individual discrimination ability of attribute 2 will produce an error of 17.08%, and the individual discrimination ability of attribute 4 will produce an error of 23.14%. When both of them are trained together, the error produced is 17.14, which is higher than the error when training

Figure 4. Architecture of neural network used to evaluate interference

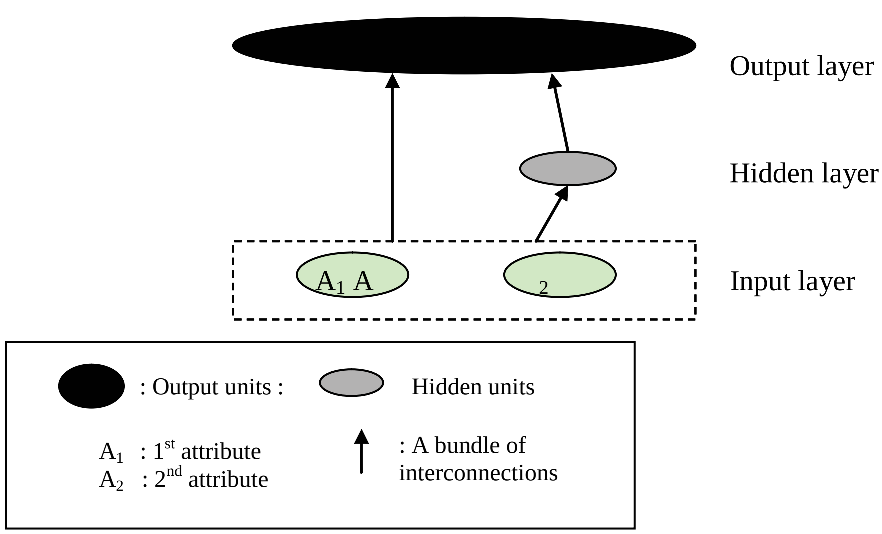


Table 1. Interference table

Attribute	1	2	3	4	5	6	……………..	N
1	IDA_1							
2	I_{12}	IDA_2						
3	I_{13}	I_{23}	IDA_3					
4	I_{14}	I_{24}	I_{34}	IDA_4				
5	I_{15}	I_{25}	I_{35}	I_{45}	IDA_5			
6	I_{16}	I_{26}	I_{36}	I_{46}	I_{56}	IDA_6		
							……………..	
N	I_{1N}	I_{2N}	I_{3N}	I_{4N}	I_{5N}	I_{6N}	……………..	IDA_N

Note : IDA_i - Individual Discrimination Ability of attribute i
I_{ij} - Classification performance of network when attribute i and attribute j are trained together

Table 2. Interference table—diabetes

% error	1	2	3	4	5	6	7	8
1	21.74							
2	16.25	17.08						
3	21.73	16.90	23.27					
4	21.64	17.14	23.09	23.14				
5	21.12	17.01	22.87	22.95	22.92			
6	20.50	16.84	21.51	21.72	21.49	21.68		
7	21.18	17.09	22.67	22.57	22.49	21.27	22.73	
8	19.61	15.04	19.85	20.05	20.11	19.37	20.20	20.02

attribute 2 alone. This means that these two attributes are interfering. The individual discrimination ability of attribute 7 will produce an error of 22.73%. When both attribute 4 and 7 are trained together, the error is 22.57, which is lower than the error produced by both the attributes' individual discriminating ability.

An overview of the Interference Table summarizing which are the interfering attributes to a particular attribute is given in Table 3. With this overview, the partitioning algorithm can proceed easily.

Interference-Free Partitioning Algorithm

The proposed interference-free partitioning method is inferred from the Interference Table. The main objective of this partitioning is to separate attributes that have interference with one another and group those attributes that are beneficial to each. The major steps involved in the algorithm are explained next.

Table 3. Overview of interference table—diabetes

Attribute	Do not interfere with	Interfere with
1	2 3 4 5 6 7 8	
2	1 3 5 6 8	4 7
3	1 2 4 5 6 7 8	
4	1 3 7	2 5 6 8
5	1 2 3 6 7	4 8
6	1 2 3 5 7 8	4
7	1 3 4 5 6	2 8
8	1 2 3 6	4 5 7

Note: The column 'Do not Interfere with' contains attributes that have no interference with the attribute in the 'Attribute' field. The column 'Interfere with' contains attributes that interfere with the attribute in the 'Attribute' field.

Partitioning Algorithm

The major steps are:

A. Assign an attribute as the default attribute, start partitioning the attributes into two batches with reference to the default attribute. One batch contains all the attributes that interfere with the default attribute, and the other batch contains those attributes that do not interfere and the default attribute itself. The assignment of the default attribute can be done randomly or in numerical order, as it is not important. Ultimately, all the attributes should be treated as the default attribute once for the whole run.

B. The next partitioning will depend on the attribute next in line to the one that has just been used for partitioning. For example, previously the partition was done using attribute 4, then the next partitioning will depend on attribute 5. If there is a total of eight attributes and the default attribute is attribute 6, and the third partitioning has been done on attribute 8, then the next partitioning will depend on attribute 1, which is next in line. Identify the batch that the reference attribute belongs to, and split the batch according to the following guideline:

   1. If none of the attributes interfere with the reference attribute within the batch, nothing is to be done. All the attributes within the batch can co-exist with each other. Also, if there is only one attribute in the batch, nothing is to be done.
   2. If the number of attributes that interfere with the reference attribute is less than or equal to the number of attributes that does not interfere within the batch, take out those interfering attributes and form a new batch. This means that within the batch, there are more attributes that can be trained with the reference attribute than those that interfere. In order to form the largest possible group, we let the reference attribute stay in the batch and take out those that interfere.
   3. If the number of attributes that interfere with the reference attribute is more than none of the interfering attributes, then take out the reference attribute from the batch.

C. Repeat Step B until the all attributes have been considered.

Steps (A) to (C) will be repeated until all attributes have been considered as the default attribute. As we try to achieve comprehensive partitioning, all attributes will take their turn to be the default attribute.

ARCHITECTURE OF IIFNNT

The network used in IIFNNT is extended from the ILIA1 and ILIA2 (Guan & Li, 2001). IIFNNT partitions the input space into different batches, and each can contain any number of attributes. Attributes within each batch are grouped together because they have no interference among each other. IIFNNT makes sure that attributes beneficial to each other are trained in the same batch, thus sharing the same subnetwork, while interfering attributes are separated and trained in different batches to reduce interference. Growing of the IIFNNT1 structure is described in the following steps:

- **Step 1.** From the grouping obtained using the interference partitioning algorithm, introduce the first batch of attributes into the neural network for training. Order of batches to be introduced into the neural network is determined by the first attribute within the batches and is done in ascending order.
  Training of the network is based on the training and validation set, which will contain only attributes that are present in the first batch. When the training is done, all the weights of the network are fixed and no longer changed.
- **Step 2.** Introduce the next batch of attributes by forming a new subnetwork. Weights and connections are only adjusted for this new subnetwork. Once the training for this new subnetwork is done, fix all the trained weights.
- **Step 3.** While there are still batches of attributes not yet introduced to the network, repeat step 2.

The IIFNNT1 is shown in Figure 5. With respect to the diabetes problem, the partitioning algorithm gives the following partition: (1 3 5 6 7) (2) (4) (8). Thus, in this case, there are four batches (i.e., four subnetworks). The first batch of attributes to be introduced into the neural network is (1 3 5 6 7) followed by a second batch that consists of attribute 2, a third batch that contains attribute 4, and last, the final batch that contains attribute 8. Figure 7 shows the IIFNNT1 for the diabetes problem.

When all the batches of attributes have been introduced into the network and the network has been fully trained, all the weights are fixed. The network is then presented with the testing set to evaluate its generalization accuracy.

Given a medical dataset for a two-class problem, there will be two output neurons in the IIFNNT architecture. The decided class of the neural networks is determined by the winner-take-all strategy. The first output neuron is assigned class 0, and the second output neuron is assigned class 1.

IIFNNT2 (Figure 6) adds more hidden units to the IIFNNT1 architecture. IIFNNT 2 hopes to obtain more information from IIFNNT1 and improve the existing network to achieve better generalization performance.

IIFNNT IN MEDICAL DIAGNOSIS

The Medical Diagnosis Datasets

The medical dataset used in our experiments is a real-world problem—the diabetes1 problem, taken from PROBEN1 benchmark collection (Prechelt, 1994). The details of the dataset are given next:

Figure 5. IIFNNT1 (ILIA1) architecture

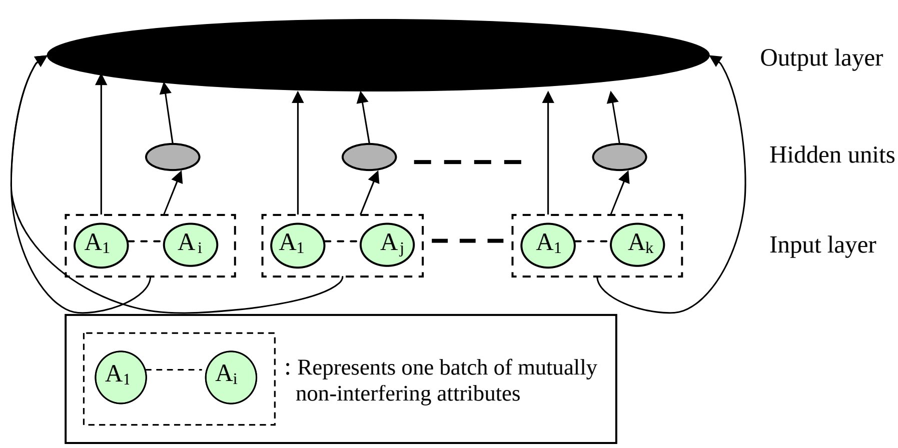


Figure 6. IIFNNT2 (ILIA2) architecture

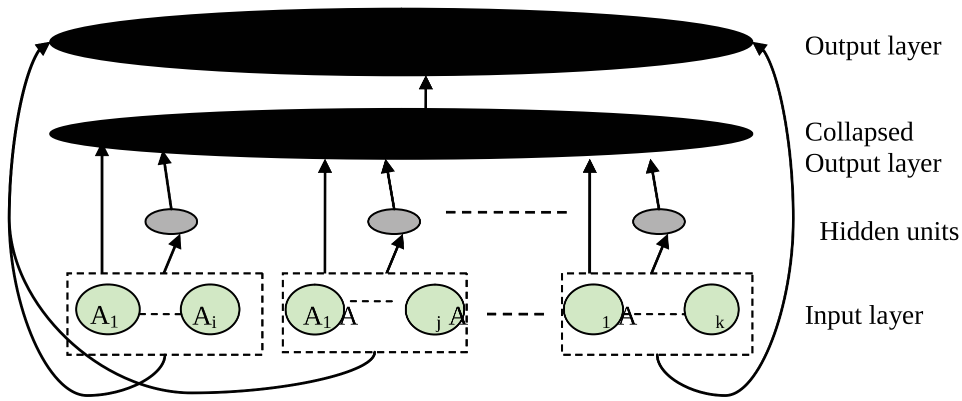
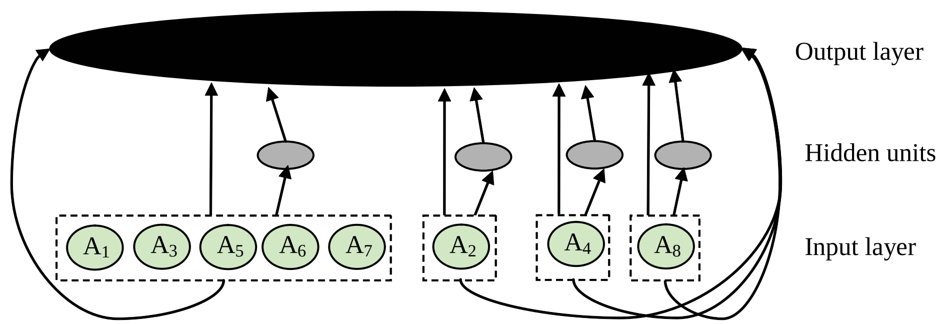


Figure 7. IIFNNT for the diabetes problem



- **Diabetes Dataset:** The "Prima Indians Diabetes" problem dataset from the UCI repository of machine learning databases was used to create this dataset. This problem diagnoses diabetes of a Prima Indian individual is based on personal data and medical examination. All the values are being encoded into continuous values in the range of 0 to 1 in PROBEN1. The class output is such that a 0 output shows no diabetes, and a 1 output shows diabetes.

Experimental Settings

All the results were taken over an average of 20 trials, and the initial assignment of weights for the neural network are from -0.25…0.25 randomly. The percentage of total examples in the dataset used for training set is 50%, validation set is 25%, and test set is 25%.

- The total number of examples used for the training set is 384; the total number of examples used for the validation set is 192; and the total number of examples used for the test set is 192.

Training of network is done using the training set; the validation set is used as a guide for stopping criteria; and the test set is used to evaluate networks generalization accuracy based on unseen data.

Experimental Results and Analysis

There are several interferences detected among the attributes of the diabetes dataset, as given in Table 2 and Table 3. The resulting partitioning by the algorithm that uses different attributes as the default attribute is shown in Table 4.

Random Ordering for Partitioning

The partitioning of the attributes by the algorithm is based on attributes being presented to it in a sequential order. If there is a high degree of interference among the input attributes, the partitioning algorithm is also able to produce different partitioning if the attributes are presented to it in a different order. Two cases of randomly ordering the attributes being presenting it to the algorithm for partitioning is investigated. Table 5 shows the diabetes dataset partitioning obtained when attributes ordering is randomized.

Best Results

Each neural network with different partitioning is being trained using the training and validation set; the resultant networks are then being evaluated for their generalization accuracy. The result from the network that gives the highest generalization accuracy is being presented in Table 7. The network with the grouping (1 2 3 6 8) (4 7) (5) gives the best results for the diabetes problem, as shown in Table 6. The standard deviation is also given in Table 7 as a guideline to the consistency of the algorithm. For the diabetes dataset, IIFNNT1 achieved better results than IIFNNT2 for the same partitioning.

Results Comparison

Our results are compared with several algorithms in literature. Not only do we compare our algorithms with neural networks used for medical dataset analysis, but we also compare them with other intelligent systems for medical diagnosis. Before we could carry out the incremental interference-free neural network training, we had to compute the Interference Table and apply the partitioning algorithm to find out the different grouping of attributes in the input space. The calculation of parameters, like the total number of hidden units and time taken for the computation of the Interference Table, is hard to materialize. Therefore, only generalization accuracy that is the major concern in this chapter is being compared.

Diabetes Dataset

The diabetes dataset results were compared with the results obtained using the conventional method of training, IFGP (Oltean & Grosan, 2003) and EFNN (Tsakonas & Dounias, 2002). The evolution of feed-forward neural networks using BNF-Grammar driven

Table 4. Partitioning obtained—diabetes

Default attribute used by algorithm	**Resulting attribute grouping**
1, 2	(1 2 3 5 6) (4 7) (8)
3, 4	(1 3 4 7) (2 5 6) (8)
5	(1 3 5 6 7) (2) (4) (8)
6	(1 3 5 6 7) (2 8) (4)
7	(1 3 4 7) (2 8) (5 6)
8	(1 2 3 6 8) (4 7) (5)

Table 5. Partitioning obtained when attributes presentation is randomized

Attributes ordering	**Attributes partitioning/grouping**
4 8 2 6 1 5 3 7	(1 3 4 7) (2 6 8) (5)
1 8 3 5 6 4 7 2	(1 2 3 6 8) (5 7) (4)

Table 6. Partitioning for diabetes dataset

Batch	Attribute #	Attribute
ONE	1	Number of times pregnant
	2	Plasma glucose concentration
	3	Diastolic blood pressure
	6	Body mass index
	8	Age
TWO	4	Triceps skin fold thickness
	7	Diabetes pedigree function
THREE	5	2-Hour serum insulin

Table 7. IIFNNT results for the diabetes dataset

Error (%) on Test set	Diabetes
Maximum (%)	23.438
Minimum (%)	21.354
Mean (%)	22.344
S.D. (%)	0.674

genetic programming (EFNN) (Tsakonas & Dounias, 2002) uses genetic programming in an attempt to automatically define feedforward neural networks.

For this problem, in terms of mean error, IIFNNT has outperformed the conventional method, IFGP and EFNN, with its low classification error of 22.344%. This is a 7%, 12%, and 3% error reduction over the conventional method, IFGP and the EFNN, respectively. Among all the algorithms mentioned previously, IIFNNT produced the smallest standard deviation. No results were presented for the maximum error and minimum error of EFNN; therefore, comparison can only be made to mean error. With our method having much better performance than the conventional method of training, it should be noted that it is important to have a partitioning algorithm in the input space to reduce the interference among input attributes. As can be seen from Table 2 and Table 3, there is a considerable amount of interference among the attributes. Maximum error produced by IIFNNT is also much lower than the maximum error (26.04%) produced by the conventional method of training the neural network. For a medical dataset, one would like to have low standard deviation in order to achieve consistent and reliable results.

FUTURE TRENDS

In the future, we will see more and more attempts to hybridize different intelligent systems to improve their overall performance in medical applications, including health care practice and delivery. For future work on the proposed approach, more research effort would be needed to find out if there exist other partitioning algorithms that could be used together with the Interference Table. It would be interesting to discover other alternative partitioning algorithms that can further improve generalization accuracy and reduce cost in medical diagnosis.

CONCLUSION

This chapter proposed input space partitioning by taking into consideration the interference among attributes to improve the accuracy in medical diagnosis. IIFNNT worked out the interference between any two attributes and separated the input data space into different

Table 8. Comparison of IIFNNT results with other algorithms—diabetes dataset

Algorithm	Mean Error (%)	Maximum Error (%)	Minimum Error (%)	Std. Dev.
IIFNNT	22.344	23.438	21.354	0.674
Conventional	23.932	26.042	21.875	1.152
IFGP	25.34	-	21.87	2.08
EFNN	23.04	-	-	-

batches, each trained with a subnetwork. By having different subnetworks for different batches, it is able to reduce the interference between the attributes. Attributes within each batch do not interfere with each other, while interfering attributes are separated. The algorithm is tested with one real world problem: the diabetes problem. The results from the experiments showed that generalization accuracy has improved with reduced interference. Neural networks are able to perform better if the correct partitioning and grouping method is applied and it increases the efficiency of network training. The cost of our method is the computation of the Interference Table. However, it should be noted that the computation of the Interference Table is a one-time cost.

REFERENCES

Abbass, H.A. (2002). An evolutionary artificial neural networks approach for breast cancer diagnosis. Artif Intell Med, 25, 265–281.

Guan, S.-U., & Li, S.C. (2001). Incremental learning with respect to new incoming input attributes. *Neural Processing Letters*, *14*(3), 241–260.

Guan, S.-U., & Li, P. (2002a). A hierarchical incremental learning approach to task decomposition. *J Intell Sys*, *12*(3), 201–226.

Guan, S.-U., & Li, P. (2002b). Feature selection for modular neural network classifiers. *J Intell Sys*, *12*(3), 173–200.

Guan, S.-U., & Li, P. (2004). Incremental learning in terms of output attributes. *J Intell Sys*, *13*(2), 95–122.

Guan, S.-U., & Liu, J. (2004). Incremental neural network training with an increasing input dimension. *J Intell Sys*, *13*(1), 43–69.

Guan, S.-U., Liu, J., & Qi, Y. (2004). An incremental approach to contribution-based attributes selection. *J Intell Sys*, *13*(1), 15–42.

Hayashi, Y., & Setiono, R. (2002). Combining neural network predictions for medical diagnosis. *Comp Bio Med*, *3*(4), 237–246.

Lehtokangas, M. (1999). Modelling with constructive backpropagation. *Neural Networks*, *12*, 707–716.

Oltean, M., & Grosan, C. (2003). *Solving classification problems using infix form genetic programming*. Proceedings of the International Symposium on Intelligence Data Analysis. Berlin: Springer-Verlag.

Plikynas, D. (2004). Decision rules extraction from neuralnNetwork: A modified pedagogical approach. *Informacines Technologijos IR Valdymas*, *31*(2), 53–59.

Prechelt, L. (1994). *PROBEN1: A set of neural network benchmark problems and benchmarking rules* [technical report 21/94]. Germany: University of Karlsruhe.

Raymer, M.L., Doom, T.E., Kuhn, L.A., & Punch, W.F. (2003). Knowledge discovery in medical and biological datasets using a hybrid bayes classifier/evolutionary algorithm. *IEEE Trans Sys, Man and Cybernetics*, *33*(5), 802–813.

Remzi, M., & Djavan, B. (2004). Artificial neural networks in urology. *European Urology Supplements*, *3*(3), 33–38.

Su, L., Guan, S.-U., & Yeo, Y.C. (2001). Incremental self-growing neural networks with the changing environment. *J Intell Sys*, *11*(1), 43–74.

Sun, R., & Peterson, T. (1999). Multi-agent reinforcement learning: Weighting and partitioning. *Neural Networks*, *12*, 727–753.

Tsakonas, A., & Dounias, G. (2002). *A scheme for the evolution of feedforward neural networks using BNF-grammar driven genetic programming*. Proceedings of the European Symposium on Intell Tech, Hybrid Sys and their Implementation on Smart Adaptive Sys, Albufeira, Portugal.

Tsakonas, A., Dounias, G., & Jantzen, J. (2004). Evolving rule-based systems in two medical domains using genetic programming. *Artif Intell Med, 32, 195–216.*

KEY TERMS

Artificial Intelligence: A branch of computer science that studies human intelligence and aims at embedding intelligent behavior, learning, and adaptation capabilities into machines.

Discrimination Ability: The ability of that particular attribute in classification performance when all the other attributes are absent from the neural network training.

Interference: Interference or cross-talking arises during the process of deciding the output value for a particular output attribute, the influence from two or more input attributes are conflicting.

Neural Network: A network of interconnecting neurons working together to produce some output function. The working of a neural network relies on the cooperation of the individual neurons within the network.

Overfitting: Refers to fitting a model (e.g., neural network) with too many samples or parameters.

Training: Refers to the process of adjusting or updating neural network weights to deliver the desired output function.

Validation: The process of checking whether overfitting (overlearning) is reached during neural network training.

Information Systems Resource Contribution in Strategic Alliance by Small Healthcare Centers

Yu-An Huang
National Chi Nan University, Taiwan

Chad Lin
Curtin University of Technology, Australia

INTRODUCTION

Alliance is defined as the "collaborative efforts between two or more firms in which the firms pool their resources in an effort to achieve mutually compatible goals that they could not achieve easily alone" (Lambe, Spekman & Hunt, 2002, p. 141). Cooperative alliances allow firms to explore new information technology, enhance a firm's knowledge foundation, lower development costs, and reduce the capital requirements and risks involved in development of new products and services (Das & Teng, 2000; Scotten, Shirin & Absher, 2006). This approach is often adopted by small and medium-sized enterprises (SMEs) as a competitive strategy to obtain necessary information systems resources in the rapidly changing and high-pressure healthcare industry (Nelson, Galvin, Essien & Levine, 1999; Scotten et al., 2006). However, according to Das and Teng (2000), around 60% of alliances between partners resulted in failure. Hence, the choice of partners and resource fit of alliance partners are of great importance for SMEs (Grant, Preece & Baetz, 1999). For those SMEs without much information systems (IS) resources, the formation process of alliance can be viewed partly as a process to increase both their tangible and intangible IS resources. Value generated from alliances is enhanced when partners have different IS resource profiles and contribute these IS resources to the alliance. These partner characteristics are important since they help in the evaluation of optimum allocations of IS resources for potential alliances to achieve suitable alliance resource alignments.

Thus, the objective of this chapter is to examine how different types of IS resource alignments affect the performance of alliances via the contribution of dissimilar and similar IS resources. The focus of this study is small healthcare centers in Taiwan. These small healthcare centers have formed alliances to confront the fierce competition as well as to absorb the regulatory pressure from the government. Most of these small healthcare centers in recent years have realized that they can only compete with major healthcare service providers through cooperative alliances, as most of the patients prefer to go to big service providers for long-term treatments. The literature on interorganizational collaborations has been criticized for its relatively narrow concentration on large firms and for ignoring SMEs' alliances where large firms do not operate in similar ways (Prater & Ghosh, 2005). These cost pressures, together with the general dynamic nature of the healthcare industry, require a significant change in approaches to utilize IS resources by these small healthcare centers. Therefore, these small healthcare centers must form alliances to obtain scarce IS resources. Unlike large firms that own a lot of IS resources to be able to form alliances with many partners, these small healthcare centers tend to form alliances with only a small number of partners, and therefore, their dependence on these partners is higher than large firms (Mambula, 2002). In this regard, these small Taiwanese healthcare centers offer an appropriate context for research.

BACKGROUND

Similar Resource Contribution

SMEs are particularly in need of finding suitable partners with dissimilar or similar resources due to their insufficient slack resources (Huang, Lin & Lin, 2005; Lin, Lin & Tsao, 2005; Srinivasan, Rangaswamy & Lilien, 2005). Formation of alliances is one way for firms that have possessed similar resources to decrease interfirm rivalry. Chen (1996) has defined resource similarity as "the degree to which two partner firms contribute resources compatible, in terms of both type and amount, to an alliance" (p. 107). The understanding of resource similarity is very important because the firms that have possessed similar resources can potentially be the fierce rivals (Chen, 1996). Moreover, resource similarity can also assist in interorganizational learning among alliance partners. For example, according to Pitts and Lei (1997), alliances formed to learn and absorb tacit knowledge are more difficult to manage among partners that come from different cultural backgrounds than partners from a similar cultural background. In addition, Inkpen (2000) has proposed that the relatedness between interpartners' knowledge base supported the acquisition of alliance knowledge. Therefore, we argue that IS resource similarity represents smaller knowledge distance and higher cultural similarity between alliance partners. These similarities are likely to assist alliance partners in developing a thorough understanding of their own knowledge and culture, and these can also enhance mutual learning and coordination and ultimately lead to increased alliance performance.

Dissimilar Resource Contribution

On the other hand, the uniquely dissimilar resource contribution is particularly critical to SMEs in their partner selection processes. SMEs are particularly in need of finding suitable partners with dissimilar resources, since it is difficult to produce alliance synergy without dissimilar resource contribution (Harrison, Hitt, Hoskisson & Ireland, 2001). Although potential competitors can easily imitate the research-based synergy from an alliance where the interpartners have strong research orientations, it will be much more difficult for them to acquire the synergy derived from dissimilar resource contribution (Harrison et al., 2001). Therefore, they are less likely to be interested in forming an alliance with firms who are only able to contribute similar IS resources. It is proposed that there is a negative correlation between interpartners on the contribution of similar IS resources. This implies that one party does not require the contribution by the other party on the IS resources they already own. On the other hand, there is a positive correlation between interpartners on the contribution of dissimilar IS resources. Interpartners that possess dissimilar IS resources are able to complement each other within an alliance. Therefore, the following hypotheses are proposed:

- **H1:** There is a negative correlation between the focal firm and the partner firm on the contribution of the similar IS resources.
- **H2:** There is a positive correlation between the focal firm and the partner firm on the contribution of the dissimilar IS resources.
- **H3:** Higher dissimilar IS resource contribution leads to higher alliance sustainable commitment.
- **H4:** Higher dissimilar IS resource contribution leads to higher alliance performance.

RESEARCH METHODOLOGY

Data Collection

In order to test the proposed hypotheses, five in-depth semistructured interviews lasting one hour each and a pilot survey of 10 industry executives were conducted to get insights into industry dynamics and to develop and refine the survey scales. Then, the main survey targeted small healthcare centers, which were classified to have formed alliances with others. This questionnaire asked the owners/directors or persons who were capable of representing their views to complete and return the questionnaire. The respondents were asked to answer the questions in relation to the cooperative relationship with their most important alliance partner.

A total of 69 questionnaires were received from the small healthcare centers that have formed alliances with others in Taiwan. The sample size is comparable to many other similar studies conducted in the last few years. For example, Sarkar, Echambadi, Cavusgil, and Aulakh (2001) collected 68 responses from 561 large international contractors sent (a net response rate of 12.1%) in their study of alliance on the role of dissimilar resource contribution, compatibility, and relationship capital on alliance performance. Nonresponse bias was tested by comparing the early and late responders on all constructs, and no significant difference between the two groups was found.

Measurement

This study has defined the interorganizational IS resource cooperation among the healthcare centers as the unique and valuable IS resources contribution by both the focal and partner firms. According to the resource-based viewpoint, IS resources (for a healthcare center) can be categorized as specialized IS support, medical IS support services, and IS management systems (Short, Palmer & Ketchen, 2002). There were 20 questions within the questionnaire, and respondents were asked to indicate their agreement on a seven-point Likert scale (1 for strongly disagree and 7 for strongly agree) with statements concerning four main constructs: (1) alliance sustainable commitment; (2) alliance performance; (3) focal firm's IS resource contribution—intangible resource, physical resource, and organizational capabilities; and (4) partner firm's IS resource contribution—intangible resource, physical resource, and organizational capabilities. In addition, we have used focal firm's perceptual sustainable commitment and performance to represent alliance sustainable commitment and performance in accordance with the finding by Geringer and Hebert (1991), in which they found that subjective performance is positively related to the objective measures of international joint venture (IJV) performance (i.e., survival, stability, and duration).

Factor analysis was performed on these measurement items to ensure that all scales were unidimensional and to assess reliability. A factor analysis was then performed to examine question items in each construct, and questions with an item loading below 0.5 were deleted without losing the representation of each of the constructs. Cronbach's alphas for all constructs are all above 0.80, indicating an acceptable reliability of the measures.

The *sustainable commitment* scale was derived from Sarkar, et al. (2001). We revised the scale into a four-item scale, and the alpha value for this scale was 0.94, indicating acceptable values of internal consistency (Nunnally, 1978). This scale measured the focal firm's willingness to invest required IS resources into the alliance.

The *performance* scale was based on Sarkar, et al.'s (2001) perceptual measure of assessing performance with both the relationship and performance between IJV partners. In this survey, performance was used as an indicator for alliance performance and measured on a seven-point Likert-type scale. Moreover, four items on healthcare characteristics were used as control variables. These included individual rating and size of healthcare centers, length of cooperative alliance, and type of cooperative institution (i.e., hospital or other healthcare service provider).

According to Johnson, Cullen, Sakano, and Takenouchi (1996), the resource contributions from both the focal and partner firms are part of the formative measurements, and as such, traditional associational-based validation procedures do not apply. Therefore, the measures for *sustainable commitment* and *performance* were analyzed for reliability and validity in accordance with the guidelines set out by Agarwal and Karahanna (2000). The reliability of the research constructs for sustainable commitment and performance was evaluated using Cronbach's coefficient alpha (α). The α values for *sustainable commitment* and *performance* in the sample were 0.94 and 0.84, respectively.

FINDINGS AND DISCUSSION

Most small healthcare centers surveyed (61%) operated independently within small and medium-sized hospitals, while the remaining small healthcare centers (39%) operated as local medical clinics. In terms of the number of hospital beds, 36% of small healthcare centers had fewer than 19 beds, while the other 39% had between 21 and 30 beds. The remaining 25% small healthcare centers had more than 30 beds. In terms of the alliance age, 38% of the respondents had less than three years, and 36% had between 3 and 5 years. The remaining 26% had more than six years.

Table 1 presents the descriptive statistics and Pearson correlation for the variables used in this study. The statistical software, SPSS 13, was used to analyze the data. Although there were few slightly stronger correlations among some independent variables, the maximum variance inflation factor (VIF) for the independent variables in all models was less than 2.766, and the average VIF was less than 1.754. This indicated that multicollinearity was not unduly influencing the least squares estimates (Neter, Wasserman & Kutner, 1985). Therefore, there was no evidence of multicollinearity in the data.

Table 2 presents the correlation matrix to highlight the relationships between the focal firms and the partner firms within interorganizational IS resources cooperation. The items shown in Table 2 are the types of resources that can be contributed by the partners. The results in Table 2 demonstrated that between the focal firm and the partner firm, there was a negative correlation on the contribution of similar IS resources. On the other hand, there was a positive correlation on

Table 1. Correlation matrix for variables

	Mean	1	2	3	4	5	6	7	8	9	10	11	12	13
Contributed by focal firms														
1. Specialized IS support	5.520	1.000												
2. Medical IS support services	4.955	0.000	1.000											
3 .IS management systems	5.561	0.000	0.000	1.000										
Contributed by partner firms														
4. Specialized IS support	4.871	-0.222^{+}	0.260^{+}	0.109	1.000									
5. Medical IS support services	4.120	0.311*	-0.524**	0.279^{+}	0.000	1.000								
6. IS management systems	4.876	0.492**	0.354*	-0.140	0.000	0.000	1.000							
Control variables														
7. Ratings	3.880	-0.211	0.116	-0.071	-0.012	-0.083	-0.112	1.000						
8. Size†	1.046	0.023	-0.102	-0.011	0.104	0.119	0.004	-0.440**	1.000					
9. Length of cooperation†	0.560	0.100	-0.023	0.240^{+}	0.075	0.158	0.112	-0.210^{+}	0.281*	1.000				
10. Hospital$^{\#}$	0.373	-0.133	-0.503**	0.017	0.095	0.361*	-0.556**	-0.197	0.115	0.032	1.000			
11. Healthcare provider$^{\#}$	0.392	0.140	0.406**	0.067	0.285*	-0.107	0.510**	0.096	0.025	-0.196	-0.619**	1.000		
12. Performance	5.672	0.119	-0.110	0.178	0.506**	0.232^{+}	-0.028	-0.298*	0.405**	0.274^{+}	0.177	0.232^{+}	1.000	
13. Commitment	6.304	0.247^{+}	0.522**	0.200	0.219	-0.015	0.417**	0.029	0.187	0.241^{+}	-0.355*	0.286*	0.000	1.000

† *Both size and length of cooperation were applied with a logarithm transformation.*
$^{\#}$ *Dummy variables.*

the contribution of dissimilar IS resources between interpartners. This supported the initial concept of similar/dissimilar IS resource contribution mentioned earlier in the chapter, since most of correlation coefficients were found to be significant. Thus, *H1* and *H2* were supported.

Models 1 and 2 in Table 3 analyzed the relationship between the individual IS resource contribution by both parties and the alliance sustainable commitment. The regression analysis results in Table 3 indicated that the contribution of specialized IS support items and medical IS support services items by the focal firms had significant positive effect on the alliance sustainable commitment. In addition, the results also showed that the contribution of IS management systems items by the partner firms had a positive impact on the alliance sustainable commitment. From these findings, it can be concluded that the contribution of dissimilar IS resources by both the focal and partner firms had a significant positive impact on the alliance sustainable commitment. Thus, *H3* was also supported.

Models 3 and 4 in Table 3 analyzed the relationship between the individual IS resource contribution by both parties and the alliance performance. The regression analysis results in Table 3 showed that the contribution of these three items (i.e., specialized IS support, medical IS support services, and IS management systems) by the focal firms had no significant relationship with the alliance performance. However, there was a positive relationship between the contribution of specialized IS support by the partner firms and the alliance performance. Therefore, as long as the partner firms were able to make significant contributions on other IS resource items, it can still have a positive effect on the alliance performance. Thus, *H4* was supported.

CONCLUSION

These findings have shown the dissimilar nature of IS resources contributed by all partners in the healthcare industry. The fact that there was a negative correlation between the similar IS resources provided by both the focal and partner firms indicates that the focal firms do not need similar IS resources provided by the partner firms. In terms of dissimilar IS resources, there was a positive correlation between the resources contributed by both the focal and partner firms. This has shown that the partner firms were able to provide necessary IS resources to satisfy needs of the focal firms. Moreover, there was a positive relationship between the IS resources contributed by both the focal and partner firms and the alliance sustainable commitment and performance. There are two approaches to establish the relationship between dissimilar IS resource contribution and alliance performance in the healthcare industry. One approach is that as long as the partner firms were able to make significant contributions on some IS resource categories, even when the focal firms failed to contribute significant unique IS resources, it could still have a positive effect on the alliance performance. The other approach to establishing a successful relationship was through both the focal and partner firms contributing unique dissimilar IS resources into the alliance. The implication of the findings is that symmetrically dissimilar IS resource contributions by interpartners within the healthcare industry can increase alliance sustainable commitment. Therefore, it can be said that successful alliance depends on partners' willingness to contribute a symmetric share of the requisite IS resources. Focal SMEs are only willing to fully commit themselves

Table 2. Correlations for the focal and partner firms' resource contribution

		Focal firms		
		Specialized IS support	Medical IS support services	IS management systems
Partner firms	Specialized IS support	-0.222^{+}	0.260^{+}	0.109
	Medical IS support services	0.311^{*}	-0.524^{**}	0.279^{+}
	IS management systems	0.492^{**}	0.354^{*}	-0.140

$^{+}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 3. The effects of IS resource contribution and performance: Unstandardized regression coefficients

	Sustainable Commitment				Performance			
	Model 1		Model 2		Model 3		Model 4	
	Beta	S.D.	Beta	S.D.	Beta	S.D.	Beta	S.D.
Constant	-2.042^{*}	1.001	-1.769	1.162	-2.196^{*}	1.067	-1.216	0.988
Contributed by focal firms								
1. Specialized IS support	0.240^{+}	0.122			0.060	0.130		
2. Medical IS support services	0.492^{**}	0.136			-0.076	0.145		
3. IS management systems	0.159	0.122			0.075	0.130		
Contributed by partner firms								
4. Specialized IS support			0.229	0.151			0.329^{*}	0.128
5. Medical IS support services			0.036	0.149			0.125	0.126
6. IS management systems			0.305^{+}	0.178			-0.165	0.151
Control variables								
7. Ratings	0.152	0.138	0.175	0.154	-0.064	0.147	-0.165	0.131
8. Size†	0.941	0.627	0.978	0.568	1.066	0.668	0.822	0.570
9. Length of cooperation†	0.873	0.522	0.605	0.580	0.923	0.556	0.872^{+}	0.483
10. Hospital#	-0.124	0.327	-0.540	0.444	0.976^{+}	0.346	0.476	0.378
11. Healthcare provider#	0.058	0.320	-0.116	0.425	1.211^{**}	0.341	0.877^{*}	0.362
F-value	4.584		2.558		3.797		5.527	
P	0.001^{***}		0.024^{*}		0.002^{*}		0.000^{***}	
Adjusted R^2	0.379		0.206		0.323		0.430	

$^{+}p<0.10$, $^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$

† *Both size and length of cooperation were applied with a logarithm transformation.*
\# *Dummy variables.*

to the alliance when they perceive fair dealing; that is, each partner has to provide equal or symmetric IS resource contributions in its alliance.

In addition, we further confirm that alliance performance could only be achieved when an interpartner contributed dissimilar IS resources to the alliance in the healthcare industry. Moreover, from the four models in Table 3, we fail to establish a positive relationship between the alliance performance and the overlapping similar IS resources contributed by both the focal and partner firms. Therefore, the results demonstrate that similar IS resource contribution had no impact on alliance performance. Furthermore, Table 3 has shown that the most characteristics of both the focal and partner firms had no direct effect on the alliance performance. Therefore, the dissimilar IS resources contribution was the key success factor for establishing alliances in the healthcare industry. Finally, the measurement proposed by this research had provided the means to evaluate both similar and dissimilar IS resources in the healthcare industry. The measurement could potentially be applied to both large firms and SMEs in other industries in the future.

FUTURE TRENDS

The future of strategic alliance in the healthcare industry will be based on the quality of the alliance relationships among partners and what they can contribute to the alliances. The results from this study suggest that SMEs should be careful in selecting their alliance partners as well as in evaluating the type and amount of dissimilar IS resources the partners are prepared to contribute to the alliance. Insufficient contribution of dissimilar IS resources to the alliance can potentially lead to failure. In addition, SMEs in the healthcare industry should prepare for the future by engaging in activities that make it a more attractive potential alliance partner. Moreover, the dedication to the growth and development of the strategic alliance can only be sustained if partners have trust and confidence in each other. Furthermore, there needs to be more than the desire to cooperate within alliances. The alliance partners should not only exchange information but also should share resources, risks, responsibilities, and rewards. Finally, the other two factors that should not be neglected are the strength of the leadership and knowledge of the healthcare industry by the alliance partners.

REFERENCES

Agarwal, R., & Karahanna, K. (2000). Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage. *MIS Quarterly, 24*(4), 665–694.

Chen, M.-J. (1996). Competitor analysis and interfirm rivalry: Toward a theoretical integration. *Academy of Management Review, 21*(1), 100–134.

Das, T.K., & Teng, B. (2000). Instabilities of strategic alliances: An internal tensions perspective. *Organization Science, 11*, 77–101.

Geringer, J.M., & Hebert, L. (1991). Measuring performance of international joint ventures. *Journal of International Business Studies, 22*(2), 249–263.

Grant, M., Preece, S.B., & Baetz, M.C. (1999). Dangers of dependence: The impact of strategic alliance use by small technology-based firms. *Journal of Small Business Management, 37*(2), 20–29.

Harrison, J.S., Hitt, M.A., Hoskisson, R.E., & Ireland, R.D. (2001). Resource complementarity in business combinations: Extending the logic to organizational alliances. *Journal of Management, 27*, 679–690.

Huang, Y., Lin, C., & Lin, H. (2005). Techno-economic effect of R&D outsourcing strategy for small and medium-sized enterprises: A resource-based viewpoint. I*nternational Journal of Innovation and Incubation, 2*(1), 1–22.

Inkpen, A.C. (2000). Learning through joint ventures: A framework of knowledge acquisition. *Journal of Management Studies, 37*(7), 1019–1043.

Johnson, J.L., Cullen, J.B., Sakano, T., & Takenouchi, H. (1996). Setting the stage for trust, and strategic integration in Japanese-U.S. cooperative alliances. *Journal of International Business Studies, 27*(5), 981–1004.

Lambe, C.J., Spekman, R.E., & Hunt, S.D. (2002). Alliance competence, resources, and alliance success: Conceptualization, measurement, and initial test. *Journal of the Academy of Marketing Science, 30*(2), 141–158.

Lin, K., Lin, C., & Tsao, H. (2005). IS/IT investment evaluation and benefit realization practices in Taiwanese SMEs. *Journal of Information Science and Technology, 2*(4), 44–71.

Mambula, C. (2002). Perceptions of SME growth constraints in Nigeria. *Journal of Small Business Management, 40*(1), 58–65.

Nelson, J.C., Galvin, G., Essien, J.D.K., & Levine, L.M. (1999). Public/private partners: Key factors in creating a strategic alliance for community health. *American Journal of Preventive Medicine, 16*(3S), 94–102.

Neter, J., Wasserman, W., & Kutner, M.H. (1985). *Applied linear statistical models*. Homewood, IL: Richard D. Irwin, Inc.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill, New York.

Pitts, R.A., & Lei, D. (1997). Building cooperative advantage: Managing strategic alliances to promote organizational learning. *Journal of World Business, 32,* 203–223.

Prater, E., & Ghosh, S. (2005). Current operational practices of U.S. small and medium-sized enterprises in Europe. *Journal of Small Business Management, 43*(2), 155–169.

Sarkar, M.B., Echambadi, R., Cavusgil, S.T., & Aulakh, P.S. (2001). The influence of complementarity, compatibility, and relationship capital on alliance performance. *Journal of the Academy of Marketing Science, 29*(4), 358–373.

Scotten, E., Shirin, B., & Absher, A.C. (2006). Creating community-based access to primary healthcare for the uninsured through strategic alliances and restructuring local health department programs. *Journal of Public Health Management and Practice, 12*(5), 446–451.

Short, J.C., Palmer, T.B., & Ketchen, D.J. (2002). Resources-based and strategic group influences on hospital performance. *Healthcare Management Review, 27*(4), 7–17.

Srinivasan, R., Rangaswamy, A., & Lilien, G..L. (2005). Turning adversity into advantage: Does proactive marketing during a recession pay off? [in press]. *International Journal of Research in Marketing*.

KEY TERMS

Alliance: The collaborative efforts between two or more firms in which the firms pool their resources in an effort to achieve mutually compatible goals that they could not achieve easily alone.

Alliance Sustainable Commitment: The maximum effort involved in maintaining and prolonging an ongoing relationship between alliance partners.

Resource Dissimilarity: The degree to which two partner firms contribute complementary resources in terms of both type and amount to an alliance.

Resource Similarity: The degree to which two partner firms contribute compatible resources in terms of both type and amount to an alliance.

SMEs: Small to medium-sized enterprises. The European Commission has defined SMEs as organizations that employ fewer than 250 people.

SPSS: A statistical and data management software package for analyzing collected questionnaire data.

Survey Research: A research method using questionnaires to obtain the required information.

Information Technology in Brain Intensive Therapy

Diego Liberati
Italian National Research Council, Italy

INTRODUCTION

In order to control a process, especially in a computer and instrumented assisted way, as in Brain Intensive Therapy (IT), a model of its behavior is needed. In order to do that, one must first be able to select the best set of a few (thus understandable and manageable) truly relevant variables.

In manufactured systems, also used in Brain IT, physics is in fact often quite known, with a manageable number of degrees of freedom; for instance, in robot kinematics, natural state variables may be (angular) positions and velocities.

Some natural systems are also easily characterized by state variables with physical meaning: for instance reservoirs (e.g., lakes), but also body organs with respect to soluble substances may be characterized by volumes, concentrations, fluxes, and gradients among compartments (Liberati & Turkheimer, 1999).

In other cases, the "natural" variables of the systems are not the best ones for identifying and control a process. A transformation of some of them may be needed; for instance, in the autonomous nervous control of the hearth system, the differential interbeat measure is the one that almost linearly interacts with baroreflex in the feedback control loop (Baselli et al., 1986).

In many cases, mainly for natural, especially neural processes, it is not easy to formulate a model based on variables whose physical meaning is a priori known; for instance, the electroencephalogram (EEG) is a very far field shielded measure of the interacting activity of billions of neurons. Many actors are thus playing, each one with many meaningful state variables, also depending on the investigated level, but with high reciprocal correlation. It would not be efficient to model every single variable for global monitoring such that it is not useful to take into account the kinetics of every single molecule of a gas when only global effects of pressure are of interest. A quest of some higher-level variables, even without a direct physical meaning, is thus natural in order to easily manage the complexity of the problem. Once such salient variables are found, the problem often arises to correlate in logical and/or mathematical sense their dynamics in order to properly model, forecast, and control the patient features.

Such interrelated problems will be addressed briefly in the present contribution, where Intensive Therapy is chosen as a paradigmatic application because of its critical conditions, while the approaches described whose rationale is better analyzed in the referred bibliography are of a quite general use in health information systems.

BACKGROUND

When data are in a time series (e.g., when a continuous process is properly sampled in time by an analog-to-digital converter at a sufficient frequency over the considered span of measured variables), a quite general approach is to express the measured sample x(t) of each variable x as a (linear) combination of:

- The recent past x(t-i) of itself;
- Possible other measured variables (which are simply taken into account in mathematical notation by considering x(t) as a vector of all the considered variables);
- A superposed noise w(t).

The EEG, for instance, is in fact well modeled by an autoregressive model of low order n:

$$x(t)=a(1)x(t-1)+a(2)x(t-2)+ \ldots +a(n)x(t-n)+w(t)$$

The set of parameters a(i), possibly varying in time, may be identified from the data time series and describes the changes in the system expressing the measured signal, whose spectrum is also easily computed from the a(i) themselves.

Such a simple black box model does allow a signal identification, useful, for instance, to monitor the effects of the control procedure; EEG may then be used in order to monitor control of drug infusion (e.g., in anesthesia) and their effects (e.g., hypotension) even in neurosurgery (Cerutti, Liberati & Mascellani, 1985) or Brain IT. Image fluxes can be treated as a set of pixel (Liberati, DiCorrado & Mandelli, 1992) or voxel (Maieron et al., 2002) streams taking into account their correlations.

In that respect, EEG, while summarizing the behavior of the complex system from which it is derived, may be quite simply modeled, processed, and identified in order to assist control. Extraction of its salient characteristics via parametric identification and/or neural networks makes the task of classifying brain states easier; even the decision to perform may be detected (Babiloni et al., 2000) toward Brain Computer Interfacing (among the 10 emerging technologies that will change the world, according to the *MIT Technical Review* of January 19, 2004). Besides being one of the most important and complex signals to be monitored in intensive therapy, and being itself the discriminant one among life and death even on legal ground, EEG thus can be reasonably assumed as the paradigmatic example of our approach to Brain IT, even involving other signals, as described in the next section.

INFORMATION PROCESSING IN BRAIN IT

When a model describing the system has been identified (e.g., in the black box way described in the background for the EEG), a direct convolution with a known input is able to render a "clean" filtered emulation of the expected output, making easier the clinical inspection. This is interesting, for instance, for single trial-evoked and event-related brain potentials (Liberati, Bedarida, Brandazza, & Cerutti, 1991a) and useful for objective assessment of possibly noncooperating subjects.

On the contrary, when only the noisy output is measurable (e.g., in hormone blood concentration), the inverse deconvolution provides an estimate of the not directly accessible variable; for instance, pituitary secretion (De Nicolao, Liberati, & Sartorio, 2000), whose little quantity as well as its little accessibility make it of unreachable direct measure.

Correlation and coherence, even partial (Liberati, Cursi, Locatelli, Comi & Cerutti, 1997), among subsystems may be usefully identified from multiple signals, discovering, for instance:

- The cardiovascular effects of the psycho-physiological stress through neural regulation (Pagani et al., 1991)
- The integrity of even nonlinear interactions between somato-sensory and visual cortex when stimulated together (Liberati et al., 1991a)
- Brain decay in degenerative pathologies (Locatelli, Cursi, Liberati, Franceschi, & Comi, 1998), or on the opposite side, hypercorrelations such as epilepsy that could, for instance, take relief by a sort of implanted defibrillator.

Time varying coefficients may be identified via Kalman filtering (Liberati, Bertolini, & Colombo, 1991b). Higher order spectral correlation also may be identified through a parametric approach, leading, for instance, to identify nonlinear effects as in muscle contraction (Orizio, Liberati, Locatelli, DeGrandis, & Veicsteinas, 1996) or in neural plasticity (Locatelli et al., 1986).

Nonetheless, linear, time invariant, and even multivariable modeling do allow much easier theory and algorithms, and thus more understandable results than nonlinear and time varying approaches. A natural quest is thus to approximate nonlinear models to linear ones in different regions. The problem is that it is not easy to a priori know-how to partition such regions. A possible solution (Ferrari-Trecate, Muselli, Liberati, & Morari, 2003) automatically clusters data in regions, without imposing continuity even in the variables at the borders, within the framework of the dynamic-logical "hybrid" models. In such a way, a linear regression is then jointly estimated in each region. Switching and nonlinear models thus may be simply identified via such Piece Wise Affine identification approach obtained by complementing a modified version of the cited stochastic parametric identification algorithm with a slightly modified *k-means* clustering. Most of the previously cited results can thus be revisited with such an approach. For instance, hormone pulses during secretion may be detected as belonging to one of the two identified states, the other being quiescent. In the same way, brain states can be discriminated, as in sleep staging.

When time course is not of paramount relevance to the data set, clustering becomes the only focus of the processing toward selecting salient variables and possibly identifying logical relationships among them. When many variables, even not homogeneous, are available, as in Brain IT, a simple hierarchy according to their interstate variance does not seem the best approach, not taking into account, besides scaling factors, correlation among them.

A Principal Components hierarchy does seem more appropriate, at least for the subset of homogeneous variables, in order to capture co-variances and to select a little number of co-factors describing a complex process, as in Garatti, Bittanti, Liberati, and Maffezzoli. (2007), where two kinds of leukemia are accurately discriminated from micro-arrays data.

A viable alternative is a Minimum Description Length hierarchy; the original variables are ranked accordingly to their decreasing ability to minimally (in terms of number of bits) codify, each alone, both the model and the observed probability distribution (Bosin, Dessi, Liberati & Pes, 2006a). Consequently, an Adaptive Bayesian Network provides a logical description of the (possibly nonlinear) interaction among them.

Such nonlinear interactions among variables also may be captured via Logical Networks (Muselli & Liberati, 2002), able to infer logical relationships among binary variables, in the form:

“if (V4 and V2) or (V3 and not V2), then....”

Their computational burden is only quadratic and, moreover, in terms of logical operations. They proved to be successful in identifying prognostic factors in oncology (Paoli, Muselli, Bellazzi, Corvó, Liberati & Foppiano, 2000). Nonlinear interrelations are detected, pointing out the relevance of the co-factor interaction, not instead provided via the widely spread, well-known powerful artificial neural networks approach (Drago, Setti, Licitra, & Liberati, 2002).

All such approaches can be fruitfully combined when identifying the salient variables among the measured ones, in order to keep directly monitoring only a little group of easily manageable size, and to identify their relationships in order also to easily understand of the underlying processes.

FUTURE TRENDS

The described approaches are not yet fully implemented everywhere, not even to some of the most critical problems in health systems such as the ones briefly introduced in the present contribution, but they are penetrating various contexts, helping to face a variety of problems even beyond those briefly addressed here for the sake of space.

Besides improvements and spreading of the introduced methodologies, a further trend should be mentioned; namely, the technological availability of Web services possibly over grid technology, able to ease cooperation among different competences even at different locations, which is now transforming forefront science as a distributed enterprise (Bosin, Dessi, Fugini, Liberati & Pes, 2006b) and will probably also penetrate health systems, as almost all useful technologies have already proved ever since.

CONCLUSION

In summary, the recalled advanced techniques of data and signal analysis, together with other more or less typical approach in health system context, may be of great help in both:

- Selecting really salient variables in discriminating classes of subjects; and
- Providing a first model describing their logical and/or dynamic interaction

Then the identified model can be used as the basis of real time processing, providing assistance also to Brain IT control, but generally useful in most applications, even beyond health systems. Interestingly, the recalled techniques, presented in the crucial framework of Brain IT, are of such a general purpose application that they are able to benefit many other applications even beyond health care.

REFERENCES

Babiloni, F., et al. (2000). Comparison between human and ANN detection of laplacian-derived electroencephalographic activity related to unilateral voluntary movements. *Comput Biomed Res, 33,* 59–74.

Baselli, G., et al. (1986). Spectral and cross-spectral analysis of heart rate and arterial blood pressure variability signals. *Comput Biomed Res, 19,* 520–534.

Bosin, A., Dessì, N., Fugini, M.G., Liberati, D., & Pes, B. (2006b). Applying enterprise models to design cooperative scientific environments. *Lecture Notes in Computer Science, 3812,* 281–292.

Bosin, A., Dessì, N., Liberati, D., & Pes, B. (2006a). Learning Bayesian classifiers from gene-expression microarray data. *Lecture Notes in Artificial Intelligence, 3849*, 297–304.

Cerutti, S., Liberati, D., & Mascellani, P. (1985). Parameters extraction in EEG processing during riskful neurosurgical operations. *Signal Processing, 9*, 25–35.

DeNicolao, G., Liberati, D., & Sartorio, A. (2000). Stimulated secretion of pituitary hormones in normal humans: A novel direct assessment from blood concentrations. *Ann Biomed Eng, 28*, 1131–1145.

Drago, G.P., Setti, E., Licitra, L., & Liberati, D. (2002). Forecasting the performance status of head and neck cancer patient treatment by an interval arithmetic pruned perceptron. *IEEE T Bio-Med Eng. 49*(8). 782–787.

Ferrari Trecate, G., Muselli, M., Liberati, D., & Morari, M. (2003). A clustering technique for the identification of piecewise affine systems. *Automatica, 39*, 205–217.

Garatti, S., Bittanti, S., Liberati, D., & Maffezzoli, P. (2007). An unsupervised clustering approach for leukemia classification based on DNA micro-arrays data. *Intelligent Data Analysis, 11*(2), 175-188.

Liberati, D., Bedarida, L., Brandazza, P., & Cerutti, S. (1991a). A model for the cortico-cortical neural interaction in multisensory evoked potentials. *IEEE T Bio-Med Eng, 38*(9), 879–890.

Liberati, D., Bertolini, L., & Colombo, D.C. (1991b). Parametric method for the detection of inter and intrasweep variability in VEP's processing. *Med Biol Eng Comput, 29*, 159–166.

Liberati, D., Cursi, M., Locatelli, T., Comi, G., & Cerutti, S. (1997). Total and partial coherence of spontaneous and evoked EEG by means of multi-variable autoregressive processing. *Med Biol Eng Comput. 35*(2), 124–130.

Liberati, D., DiCorrado, S., & Mandelli, S. (1992). Topographic mapping of single-sweep evoked potentials in the brain. *IEEE T Bio-Med Eng, 39*(9), 943–951.

Liberati, D., & Turkheimer, F. (1999). Linear spectral deconvolution of catabolic plasma concentration decay in dialysis. *Med Biol Eng Comput, 37*, 391–395.

Locatelli, C., et al. (1986). High order frequency domain analysis of population activity in rat auditory cortex. *Journal of Physiology-London, 493P*(S20–S21).

Locatelli, T., Cursi, M., Liberati, D., Franceschi, M., & Comi, G. (1998). EEG coherence in Alzheimer's disease. *Electroen Clin Neuro, 106*(3), 229–237.

Maieron, M., et al. (2002). An ARX model-based approach to the analysis of single-trial event-related fMRI data. In *Proceedings of the 8th International Conference on Functional Mapping of the Human Brain*, Sendai, Japan.

Muselli, M., & Liberati, D. (2002). Binary rule generation via hamming clustering. *IEEE T Knowl Data En, 14*(6), 1258–1268.

Orizio, C., Liberati, D., Locatelli, C., DeGrandis, D., & Veicsteinas, A. (1996). Surface mechanomyogram reflects muscle fibres twitches summation. *J Biomech, 29*(4), 475–481.

Pagani, M., et al. (1991). Sympatho-vagal interaction during mental stress: A study employing spectral analysis of heart rate variability in healthy controls and in patients with a prior myocardial infarction. *Circulation, 83*(4)II, 43–51.

Paoli, G., Muselli, M., Bellazzi, R., Corvó, R., Liberati, D., & Foppiano, F. (2000). Hamming clustering techniques for the identification of prognostic indices in patients with advanced head and neck cancer treated with radiation therapy. *Med Biol Eng Comput, 38*, 483–486.

KEY TERMS

Adaptive Bayesian Networks: Conditional tree of probabilistic dependence among features discriminating salient classes of data.

Autoregressive Model: Simple description of the time course of a variable in terms of the recent past of the variable itself.

Brain Computer Interfacing: Direct command of computers or instruments via brain waves processing without need of usual actuators (e.g., fingers and keypads, or voice and microphones).

Clustering: Partition of a set of data in groups (clusters) in such a way that each datum is maximally close in some defined metric to all the others belonging to the same cluster and maximally far from every other datum belonging to different clusters.

Coherence: Phase synchrony among pairs of variables, even with a fixed delay.

Evoked Brain Potentials: Brain activity directly related to the execution of a peculiar task.

Information Technology: The set of approaches related to the processing of information through both electronic hardware and mathematical software.

Intensive Therapy: One of the most demanding divisions in health care, where a patient's life is at risk and then where ultimate approaches (e.g., the ones presented in this chapter) may be more effective, even if they are also useful in other contexts beyond health systems.

Piece-Wise Affine Identification: Decomposition of a multivariable sampling of a possible nonlinearly related and/or time varying set of variables in linearly related sections, with automatic identification of the optimal borders among sections through joint clustering and autoregressive identification.

Innovative Piezoelectric Extracorporeal Lithotripter

Achim M. Loske
Universidad Nacional Autónoma de México, Mexico

Francisco Fernández
Universidad Nacional Autónoma de México, Mexico

Gilberto Fernández
Universidad Nacional Autónoma de México, Mexico

INTRODUCTION

Before 1980, the majority of patients with urolithiasis and nephrolithiasis needed surgery (Kerbl, Rehman, Landman, Lee, Sundaram, & Clayman, 2002; Soucie et al., 1994). Fortunately, percutaneous nephrolithotomy, ureteroscopic intrarenal surgery, laparoscopic surgery, and extracorporeal shock wave lithotripsy (SWL) now allow almost any calculus to be removed without open surgery. SWL refers to the use of high intensity pressure pulses, generated outside the body, to break up kidney stones (Chaussy, Brendel, & Schmiedt, 1980; Loske, 2007). It has become the standard treatment for the majority of patients and an alternative in the management of gallbladder stones, pancreatic concrements, and salivary gland stones. Even though initial studies concluded that shock waves had no damaging effect on renal tissue, later several authors reported that shock waves may cause tissue trauma (Evan, Willis, Connors, McAteer, & Lingeman, 1991; Evan, Willis, & Lingeman, 2003 Willis et al., 1999). Fortunately, techniques and devices are still evolving and improvements to increase stone fragmentation efficiency and reduce tissue trauma are being constantly sought.

BACKGROUND

The apparatus to perform SWL, called lithotripters, are composed of a shock wave generator, a focusing unit, an imaging system, a coupling device, and a treatment table (Lingeman, 2007; Loske, 2007). If the patient is properly positioned, shock waves enter the body, become focused on the calculus, and fracture it. Shock wave targeting is performed by fluoroscopy or ultrasound imaging systems. Several hundred shock waves may be needed to comminute a kidney stone. Fragments pass spontaneously through the urinary tract, and most patients are free of stones a few days after treatment. Three shock wave generation techniques are in use: piezoelectric, electrohydraulic, and electromagnetic (Loske, 2007). Many lithotripters use a focusing system to concentrate the shock wave energy onto the stone; others are self-focusing. Only piezoelectric shock wave generation will be discussed here.

Piezoelectric lithotripters generate shock waves by a high voltage discharge applied across an array of up to 3000 piezoelectric crystals mounted on a hemispherical bowl-shaped aluminum backing (Figure 1). Other systems only use a few crystals or one large spherically shaped ceramic plate. Each high voltage discharge causes rapid expansion of each crystal, producing a pressure wave. The pressure pulse arriving at the center (F) is generated by superposition of the pulse produced by each crystal (Loske, 2007). The shock wave generator is contained inside a cavity filled with water and closed with a rubber membrane placed in contact with the skin of the patient. Piezoelectric crystals are insulated from water with a flexible polymer. These lithotripters produce a focal zone in the shape of a cigar measuring about 17 × 3 mm. The shock wave pulse rate and the discharge voltage can be varied. The pressure in the focal region (Figure 2) consists of a compression pulse with a peak between 30 and 150 MPa and a phase duration of 0.5 to 3 μs, followed by a decompression pulse, sometimes referred to as the "negative" pressure pulse, with a tensile peak of up to - 20 MPa, and a phase duration between 2 and 20 μs. Pressure rise time is about 300 ns.

Cavitation is one of the main stone comminution mechanisms. Calculi also fracture due to spalling, squeezing, superfocusing, and fatigue (Eisenmenger,

Figure 1. Schematic of a standard piezoelectric shock wave generator. The power supply P charges a capacitor C, which remains charged until the switch S is fired by the trigger T. Shock wave rate can be adjusted with the pulse generator G.

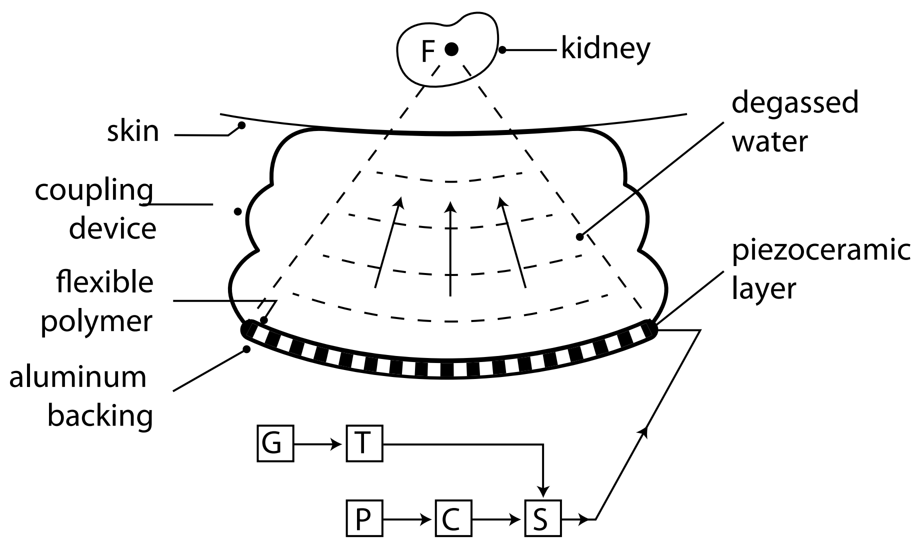
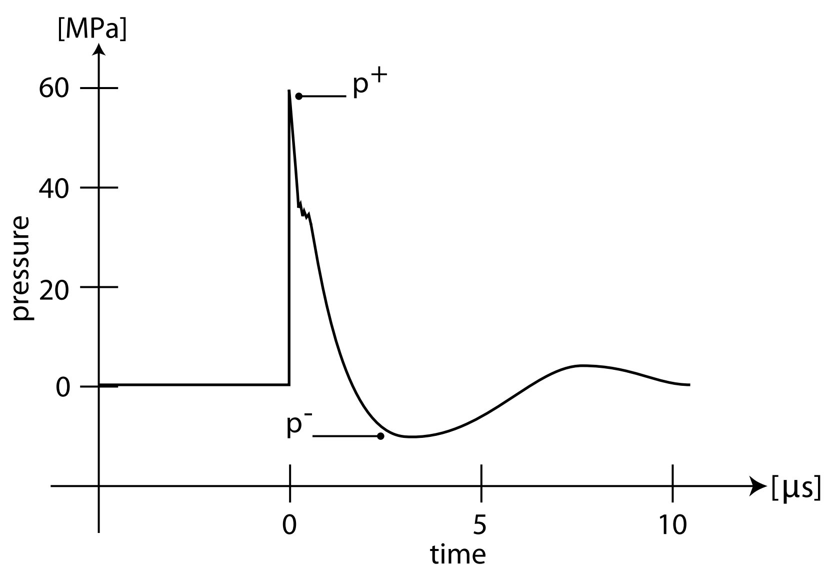


Figure 2. Pressure record obtained with a fiber optic hydrophone (FOPH 2000, RP Acoustics, Leutenbach, Germany), positioned at the focus of a standard piezoelectric lithotripter.



2001; Lokhandwalla, & Sturtevant, 2000; Loske, 2007). During SWL, all microbubbles contained in the fluid near the calculus are compressed by the positive peak (Figure 3). Their volume increases about 50 to 100 µs after shock wave passage. This occurs due to the action of the tensile part of the shock wave (Bailey et al., 2005). Lithotripter generated bubbles expand, stabilize, and collapse violently after approximately 250 to 500 µs, producing high-speed microjets (Crum, 1988), capable of damaging the calculus.

Enhanced fragmentation can be obtained if a second shock wave arrives during or shortly after their stable phase (Figure 4). This has been demonstrated with twin pulses generated from two identical shock

Figure 3. Representation of the pressure variation at the focus of a standard lithotripter and its influence on a bubble. The bubble is compressed as the positive pulse arrives (0 µs). Its volume increases a few microseconds later due to the tensile phase of the shock wave, stabilizes, and finally collapses after about 330 µs.

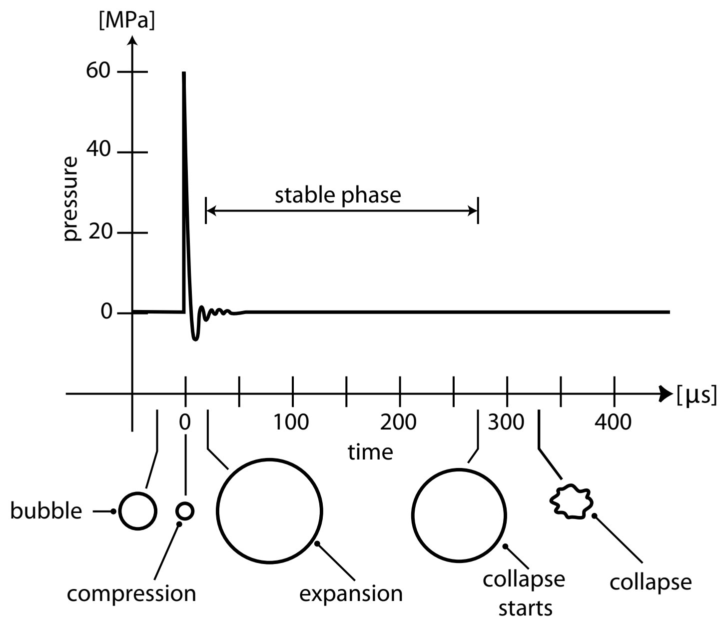
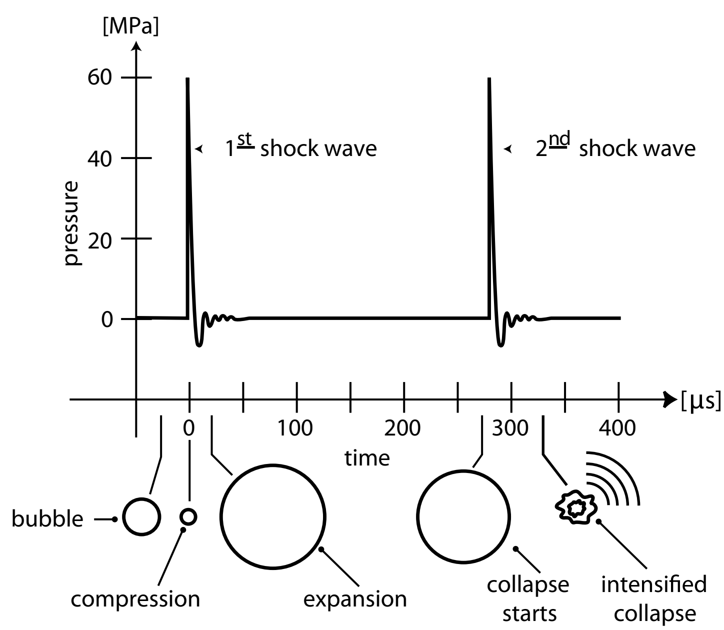


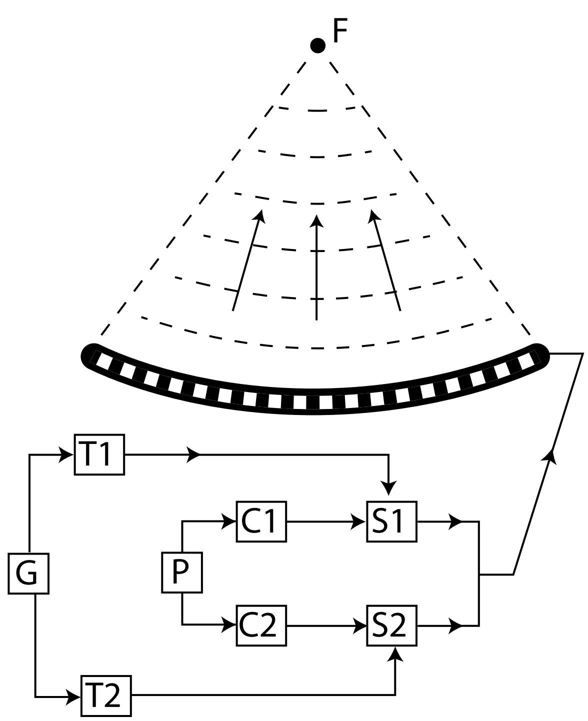
Figure 4. Representation of the pressure variation at the focus of the novel lithotripter and its influence on a bubble. The collapse of a bubble expanded by the first shock wave can be intensified if a second shock wave arrives shortly after the collapse starts.



wave sources or using electrohydraulic shock wave generators with hybrid-ellipsoidal reflectors (Bailey, 1997; Loske & Prieto, 1996; Loske, Prieto, Gutiérrez, Zendejas, Saita, & Vélez, 2004; Prieto & Loske, 1999; Sheir, El-Diasty, & Ismail, 2005; Sokolov, Bailey, & Crum, 2000; Sokolov, Bailey, & Crum, 2001; Zhong et al., 1997; Zhong et al., 1999). Other authors (Xi & Zhong, 2000) proposed a combined electrohydraulic and piezoelectric shock wave generator.

Disadvantages of the above-mentioned systems are the need for two shock wave generators or the construction of a different composite reflector for each desired time delay between the first and the second shock wave. As a solution to this, we developed a novel piezoelectric lithotripter that generates two successive shock waves with an adjustable time delay between them (Loske, Fernández, & Gutiérrez, 2005; Loske, Prieto, Fernández, & van Cauwelaert, 2002). We named the device "tandem" shock wave generator, because of the extremely short delay between first and second shock wave.

Figure 5. Schematic of the new piezoelectric shock wave generator. The power supply P charges two capacitors. C1 remains charged until the switch S1 is fired by the trigger T1. C2 is discharged through S2 after a delay. The shock wave rate and the delay may be adjusted with the pulse generator G.



THE NOVEL LITHOTRIPTER

The first successful experiments with the tandem shock wave generator were performed during 2000 and 2001 (Fernández, Loske, van Cauwelaert, & Prieto, 2001; Fernández, Loske, van Cauwelaert, & Prieto, 2004; Loske, Prieto, van Cauwelaert, & Fernandez, 2001). A *Piezolith 2300* lithotripter (Richard Wolf GmbH) was converted into a tandem lithotripter by duplicating its energy discharge system (Figure 5). Both high voltage outputs were connected in parallel to the piezoelectric crystal array. A specially designed pulse generator triggers both systems with an adjustable time delay between 50 and 950 μs. The delay can be varied in steps of 10 μs.

Stone comminution obtained with the standard lithotripter was compared in vitro with that produced by the novel lithotripter at delay times between 100 and 900 μs using rectangular and spherical artificial kidney stones (High Medical Technologies). The delay was increased in steps of 50 μs. Stones were placed, one by one, inside a small polyethylene bag, and centered at F using the imaging system of the lithotripter. Six rectangular stones were exposed to 1200 standard shock waves and the same number of stones was treated with 600 tandem pulses. Additionally, a set of 6 spherical stones received 1000 standard, and an equal set received 500 tandem shock waves. Fragments were strained through a 1.0 x 1.0 mm nylon mesh after shock wave exposure, dried in an oven, and weighed on a precision scale. The discharge frequency of the system was fixed at 1 Hz. Results were encouraging (Figure 6). The best fragmentation coefficient (FC) was obtained at a delay of about 240 and 400 μs with spherical and rectangular stones, respectively.

In vivo tissue damage to rabbit kidneys obtained with the tandem lithotripter was compared with that of the standard lithotripter. Our results revealed that the tandem shock wave generator did not produce more tissue damage than the standard system (Loske, Fernández, Zendejas, Paredes, & Castaño-Tostado, 2005). Considering that less shock waves are needed when using the novel lithotripter, SWL tissue damage could be reduced.

To study the influence of cavitation during tandem lithotripsy, 7-mm long, 5-mm in diameter cylindrical artificial kidney stones, obtained from rectangular AST stones (High Medical Technologies), were implanted into the renal pelvis of 180 g pig kidneys, and exposed

Figure 6. In vitro fragmentation coefficient obtained with spherical (FC-S), and rectangular (FC-R) kidney stone models, exposed to tandem shock waves at different delays. The units on both vertical axes differ, because the rectangular stones are much larger than the spherical stones. Nonparametric smoothing was used to obtain both graphs.

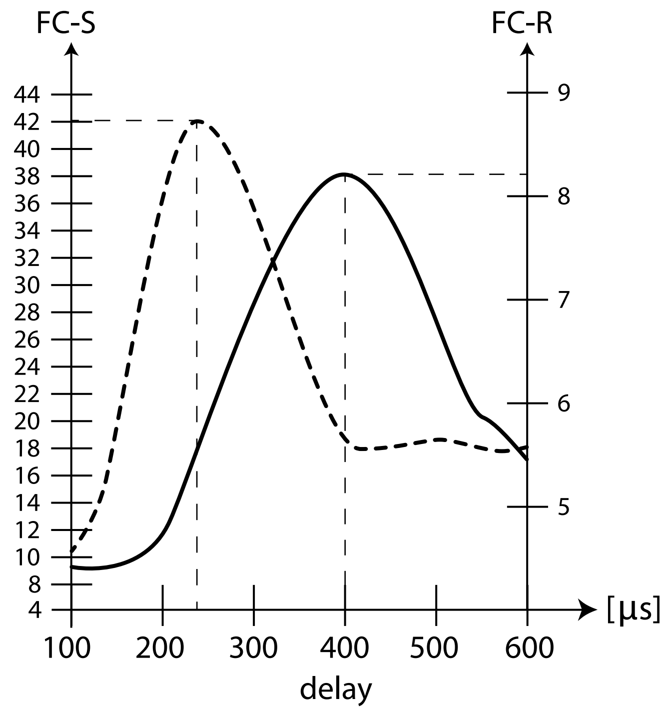
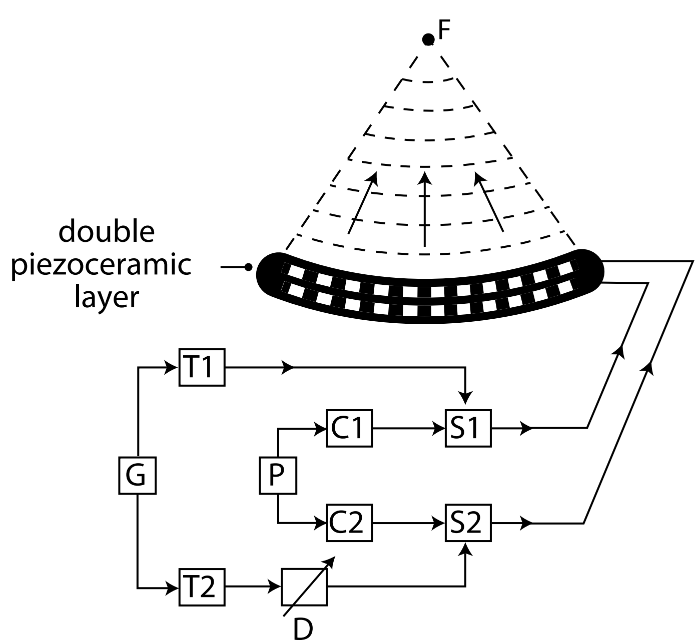


Figure 7. The innovation presented here has been used in the Piezolith 3000 and the Piezoson 100 plus (Richard Wolf GmbH). Two piezoceramic layers are activated through separate circuits. Pulses generated by both layers superpose additively, because the crystals on the inner side are fired with a delay D after the outer layer. Tandem shock waves can be generated by increasing the delay a few hundred microseconds.



in vitro to 200 (100) single (tandem) shock waves. Delays between shock waves in the tandem mode were 250, 300, 350, and 400 μs. All stones were hydrated and placed at F using the ultrasound scanners. The kidneys were placed on a specially designed holder, inside the water tank of the lithotripter. In the first part, all stones were directly inserted into the kidney. In the second part, each stone was placed inside the tip of a finger of a latex glove, filled with saline solution, so that each stone was surrounded by about one millimeter of fluid, providing an "expansion chamber." The diameter of these balloons was about 8 mm. Each balloon was inserted inside the renal pelvis of a kidney. Again, single and tandem shock waves were used (7.5 kV, 1 Hz). The whole experiment was repeated five times. As shown in Table 1, tandem shock waves did not significantly improve fragmentation efficiency in those cases where the stone was inserted into the renal pelvis without an expansion chamber; however, the novel system enhanced stone fragmentation in stones surrounded by saline solution. The existence of an expansion chamber increased the FC by about 20% when using the lithotripter in the standard mode, and 34% with the tandem mode (delay = 350 μs).

Our results demonstrate two important issues: (a) acoustic cavitation significantly contributes to stone fragmentation during SWL, and (b) tandem shock wave lithotripsy is more efficient than standard single pulse lithotripsy only if the stone is surrounded by fluid. The optimal delay for tandem shock wave lithotripsy is between 250 and 400 μs; however this will depend on the geometry, size, composition, and location of the stone.

Table 1. In vitro fragmentation coefficients (FC) of artificial kidney stones exposed to 200 standard single-pulse (delay = 0 μs), or 100 tandem-pulse shock waves inside the renal pelvis of pig kidneys. FC_1 corresponds to stones implanted into the kidney without an expansion chamber, and FC_2 to stones placed into the kidney inside a saline solution-filled latex balloon.

Delay [μs]	FC_1 [%]	FC_2 [%]
0	20.2	40.6
250	17.9	47.4
300	15.4	54.3
350	21.9	56.1
400	14.3	54.7

A concern about the new lithotripter could be increased tissue damage. Fortunately, most mammalian tissue has few cavitation nuclei (Carstensen, Gracewski, & Daleki, 2000). Furthermore, cavitation bubbles in tissue have a different behavior from that observed in liquid. Since bubble expansion in vivo is constrained by the tissue, collapse is less violent than in urine.

FUTURE TRENDS

Shock wave generators having two layers of crystals mounted on a spherical dish are already on the market (Riedlinger, Dreyer, & Krauss, 2002). In order to compensate for the difference in path to the focus, the piezoelectric crystals on the convex side are fired before those on the concave (radiating) side. Both waves superimpose at the surface of the front layer (Figure 7). In this case, the modification required to produce our tandem shock waves is easier, since these piezoelectric-crystal arrays are already excited with delay.

All piezoelectric lithotripters can be converted into tandem lithotripters at relatively low cost. Our innovation (Loske et al., 2002) has already been used by other authors (Arora, Junge, & Ohl, 2005) and was recently installed into the Piezolith 3000 and the Piezoson 100 plus (Richard Wolf GmbH, Germany) shock wave generators (Ginter & Krauss, 2005). Since preliminary clinical results did not improve with the double layer technology (Ng et al., 2003), the manufacturer expects better results using our tandem system. No clinical results have been reported so far.

Due to its flexibility, appropriate tandem shock wave delay may also be useful to improve the efficiency of shock wave mediated macromolecule delivery (Armenta, Varela, Martínez de la Escalera, & Loske, 2006) and exposure of isolated cells and microorganisms (Álvarez et al., 2004).

CONCLUSION

Shock wave-induced cavitation is one of the main fragmentation mechanisms during SWL; however, it will only be efficient if the urinary stone is surrounded by fluid. Because of this, urologists are encouraged to try to produce an expansion chamber for the calculus before starting SWL. Furthermore, if the urinary stone is surrounded by saline solution or urine, enhanced fragmentation can be obtained using the novel tandem lithotripter described here. Tandem shock wave sources could be installed in clinical devices to improve fragmentation and reduce the number of shock waves used per treatment.

ACKNOWLEDGMENT

The authors acknowledge assistance by Concepción Arredondo, Alejandra Domínguez, Arturo Méndez, and René Preza. They would also like to thank Helmuth Busch†, Eduardo Castaño, and Mario Granizo for advice. Figures were designed by Gabriela Trucco.

REFERENCES

Álvarez, U.M., Loske, A.M., Castaño-Tostado, E., & Prieto, F.E. (2004). Inactivation of *Escherichia coli* O157:H7, *Salmonella* Typhimurium and *Listeria monocytogenes* by underwater shock waves. *Innovative Food Science and Emerging Technologies, 5*(4), 459-463.

Armenta, E., Varela, A., Martínez de la Escalera, G., & Loske, A.M. (2006). Transfección de la línea celular Hela derivada de un tumor cervico-uterino por medio de ondas de choque in vitro. *Revista Mexicana de Física, 52*(4), 352-358.

Arora, M., Junge, L., & Ohl, C.D. (2005). Cavitation cluster dynamics in shock-wave lithotripsy. Part 1: Free field. *Ultrasound in Medicine and Biology, 31*(6), 827-839.

Bailey, M.R. (1997). *Control of acoustic cavitation with application to lithotripsy*. Doctoral dissertation, University of Texas at Austin (p. 200).

Bailey, M.R., Pishchalnikov, Y.A., Sapozhnikov, O.A., Cleveland, R.O., McAteer, J.A., Miller, N.A., et al. (2005). Cavitation detection during shock-wave lithotripsy. *Ultrasound in Medicine and Biology, 31*(9), 1245-1256.

Carstensen, E.L., Gracewski, S., & Daleki, D. (2000). The search for cavitation in vivo. *Ultrasound in Medicine and Biology, 26*, 1377-1385.

Chaussy, C., Brendel, W., & Schmiedt, E. (1980). Extracorporeally induced destruction of kidney stones by shock waves. *Lancet, 2,* 1265-1268.

Crum, L.A. (1988). Cavitation microjets as a contributory mechanism for renal calculi disintegration in ESWL. *The Journal of Urology, 140,* 1587-1590.

Eisenmenger, W. (2001). The mechanism of stone fragmentation in ESWL. *Ultrasound in Medicine and Biology, 27,* 683-693.

Evan, A.P., Willis, L.R., Connors, B.A., McAteer, J.A., & Lingeman, J.E. (1991). Renal injury by extracorporeal shock wave lithotripsy. *The Journal of Endourology, 5,* 25-35.

Evan, A.P., Willis, L.R., & Lingeman, J.E. (2003). Shock wave lithotripsy (SWL) induces significant structural and functional changes in the kidney. *Journal of the Acoustical Society of America, 114*(4), 2454.

Fernández, F., Loske, A.M., van Cauwelaert, J., & Prieto, F.E. (2001). Sistema electrónico para mejorar la eficiencia de litotriptores extracorporales piezoeléctricos. Mexican Society of Instrumentation. In *Proceedings of the 16th National Congress on Electronic Instrumentation*, MAT 1-1.

Fernández, F., Loske, A.M., van Cauwelaert, J., & Prieto, F.E. (2004). Electronic device to improve the efficiency of extracorporeal lithotripters. *Journal of Applied Research and Technology CCADET-UNAM, 2*(2), 170-178.

Ginter, S., & Krauss, W. (2005) Wolf: Innovative Piezoelectric shock wave systems – Piezolith 3000 and Piezoson 100 plus. In Ch. Chaussy, G. Jocham, K.U. Kohrmann & D. Wilbert (Eds.), *Therapeutic energy applications in urology. Standards and recent developments* (pp. 175-177). Georg Thieme Verlag KG.

Kerbl, K., Rehman, J., Landman, J., Lee, D., Sundaram, C., & Clayman, R.V. (2002). Current management of urolithiasis: Progress or regress? *The Journal of Endourology, 16,* 281-288.

Lingeman, J.E. (2007). Lithotripsy systems. In A.D. Smith, G.H. Badlani, D.H. Bagley, R.V. Clayman, S.G. Docimo, G.H. Jordan et al. (Eds.), *Smith's textbook on endourology* (pp. 333-342). Hamilton, Ontario, Canada: BC Decker, Inc.

Lokhandwalla, M., & Sturtevant, B. (2000). Fracture mechanics model of stone comminution in ESWL and implications for tissue damage. *Physics in Medicine and Biology, 45,* 1923-1949.

Loske, A.M. (2007). *Shock wave lithotripsy physics for urologists*. Juriquilla, Qro., Mexico: Centro de Física Aplicada y Tecnología Avanzada, UNAM. ISBN 978-970-32-4377-8.

Loske, A.M., Fernández, F., & Gutiérrez, J. (2005). Steigerung der Fragmentationseffizienz durch Verstärkung von Kavitation zur berührungsfreien Nierensteinzertrümmerung. *Zeitschrift für Medizinische Physik, 15,* 53-58.

Loske, A.M., Fernández, F., Zendejas, H., Paredes, M.I., & Castaño-Tostado, E. (2005). Dual-pulse shock wave lithotripsy: *In vitro* and *in vivo* study. *The Journal of Urology, 174*(12), 2388-2392.

Loske, A.M., & Prieto, F.E. (1996). Improving underwater shock wave focusing efficiency. In C.Y.C. Pak, M.I. Resnick & G.M. Preminger (Eds.), *Urolithiasis* (pp. 401-402). Dallas, TX: Millet The Printer.

Loske, A.M., Prieto, F.E., Fernández, F., & van Cauwelaert, J. (2002). Tandem shock wave cavitation enhancement for extracorporeal lithotripsy. *Journal of Physics in Medicine and Biology, 47*(22), 3945-3957.

Loske, A.M., Prieto, F.E., Gutiérrez, J., Zendejas, H., Saita, A., & Vélez, E. (2004). Evaluation of a bifocal reflector on a clinical lithotripter. *The Journal of Endourology, 18*(1), 7-16.

Loske, A.M., Prieto, F.E., van Cauwelaert, J., & Fernández, F. (2001). Alternative pressure waveforms for ESWL. In *Proceedings of the 17th International Congress on Acoustics*, Rivista Italiana di Acustica (Vol. 25, pp. 1-3).

Ng, C., McLoran, L., Thompson, T. J., Anagnostou, T., Houssa, & S., Smith, et al. (2003). Comparison of two generations piezoelectric lithotripers by using match-pair analysis. *Journal of Endourology, 17*(1).

Prieto, F.E., & Loske, A.M. (1999). Bifocal reflector for electrohydraulic lithotripters. *The Journal of Endourology, 13,* 65-75.

Riedlinger, R., Dreyer, T., & Krauss, W. (2002). Small aperture piezo sources for lithotripsy. In A. Bettucci (Ed.), *Proceedings of the 17th Int. Congress on Acoustics,* Rome, Italy, (Vol. IV).

Sheir, K.Z., El-Diasty, T.A., & Ismail, A.M. (2005). Evaluation of a synchronous twin-pulse technique for shock wave lithotripsy: The first prospective clinical study. *British Journal of Urology International, 95*(3), 389-393.

Sokolov, D.L., Bailey, M.R., & Crum, L.A. (2000). Effect of dual-reflector lithotripsy on stone fragmentation and cell damage. *Journal of the Acoustical Society of America, 108*(5), 2518.

Sokolov, D.L., Bailey, M.R., & Crum, L.A. (2001). Use of a dual-pulse lithotripter to generate a localized and intensified cavitation field. *Journal of the Acoustical Society of America, 110,* 1685-1695.

Soucie, J.M., Thun, M.J., Coates, R.J., McCellan, W., & Austin, H. (1994). Demographic and geographic variability of kidney stones in the United States. *Kidney International, 46*(3), 893-899.

Willis, L.R., Evan, A.P., Connors, B.A., Blomgren, P., Fineberg, N.S., & Lingeman, J.E. (1999). Relationship between kidney size renal injury, and renal impairment induced by shock wave lithotripsy. *Journal of the American Society of Nephrology, 10,* 1753-1762.

Xi, X., & Zhong, P. (2000). Improvement of stone fragmentation during shock-wave lithotripsy using a combined EH/PEAA shock-wave generator—In vitro experiments. *Ultrasound in Medicine and Biology, 26,* 457-467.

Zhong, P., Cocks, F.H., Cioanta, I., & Preminger, G.M. (1997). Controlled, forced collapse of cavitation bubbles for improved stone fragmentation during shock wave lithotripsy. *The Journal of Urology, 158*, 2323-2328.

Zhong, P., Lin, H.F., Xi, X.F., Zhu, S.L., & Bhogte, E.S. (1999). Shock wave-inertial microbubble interaction: Methodology, physical characterization, and bioeffect study. *Journal of the Acoustical Society of America, 105,* 1997-2009.

KEY TERMS

Acoustic Cavitation: Growth and collapse of vapor bubbles because of a fast change of positive pressure into tensile stress, produced by a shock wave.

Dual-Pulse Shock Waves: Shock waves generated with a dual-head lithotripter. Dual-pulse shock waves are not necessarily equivalent to tandem shock waves.

Focal Zone: Volume in which the pressure at any point is equal to or higher than 50% of the maximum peak positive pressure.

Fragmentation Coefficient: For in vitro kidney stone fragmentation, the fragmentation coefficient is defined as $FC = (W_i - W_f)100/W_i$. In this equation, W_i and W_f stand for the weights of the initial (intact) stone and final stone fragments.

Lithotripter: A device to perform SWL, composed of a shock wave generator, a focusing unit, an imaging system, a coupling device, and a patient treatment table.

Peak Negative Pressure: Maximum pressure amplitude of the negative pulse of a shock wave, that is, the difference between the minimum pressure and the ambient pressure.

Peak Positive Pressure: Maximum pressure amplitude of the positive pulse of a shock wave, that is, the difference between the maximum pressure and the ambient pressure.

Rise Time: Time needed for the pressure peak to rise from 10% to 90% of its final value.

Shock Wave: Mechanical wave of high amplitude that arises when any elastic medium is subjected to a rapid compression.

Tandem Shock Wave: Consecutive shock waves having a delay of up to several hundred microseconds, used in SWL to enhance cavitation bubble collapse and, as a result, increase fragmentation efficiency.

Insulin Metabolism Models for Children with Type 1 Diabetes

Stavroula G. Mougiakakou
National Technical University of Athens, Greece

Aikaterini Prountzou
National Technical University of Athens, Greece

Dimitra Iliopoulou
National Technical University of Athens, Greece

Andriani Vazeou
National Technical University of Athens, Greece

Christos S. Bartsocas
"MITERA" Hospital, Greece

Konstantina S. Nikita
National Technical University of Athens, Greece

INTRODUCTION

Diabetes Mellitus (DM) is a chronic metabolic disease resulted from insufficient secretion of hormone insulin. DM is mainly classified into Type 1 (or insulin dependent diabetes), which is characterized by absence of insulin secretion, due to destruction of the β-cells of pancreas, and Type 2 (or insulin independent diabetes), which is characterized by reduced action of insulin. This dysfunction of insulin results in many short- (hypoglycaemia, hyperglycaemia) and long-term (like neuropathies, nephropathies, retinopathies, and so on) complications. The result of the Diabetes Control and Complications Trial (DCCT) (DCCT Research Group, 1993) and the U.K. Prospective Diabetes Study (UKPDS) (UKPDS Group, 1998) indicate that the intensive glucaemic control reduces many short- and long-term complications of DM.

The ultimate goal in management of DM is the development of an automatic "closed-loop" system, well known as "artificial pancreas," able to simulate accurately the glucose-insulin metabolism and maintain glucose levels of a diabetes patient into physiological levels (70–110 mg/dl for a healthy person). A "closed-loop" system for DM comprises three primary components: (i) an accurate frequent or continuous glucose measurement system, (ii) an insulin delivery system, and (iii) a control system able to change the dose of delivered insulin, with respect to the requirement of glucose levels in desired range. Currently, individuals with Type 1 Diabetes Mellitus (T1DM) measure their glucose levels using either conventional finger-stick glucose meters (three to four times per day), or continuous glucose monitoring systems (CGMS), and choose their insulin delivery method between multiple daily injection (MDI), and continuous subcutaneous insulin infusion (CSII). The most appropriate algorithms to close the loop seems to be those based on model predictive control—MPC (Bequette, 2005). In a MPC controller, a model is used for the prediction of current and future insulin delivery parameters based on estimation on future glucose concentrations, while an optimizer finds the optimum insulin delivery parameter values in order to maintain future glucose values inside the desired range.

The models for the predictions of glucose values simulate the glucose-insulin metabolic system, which is characterized by high complexity and nonlinearity. Several mathematical models (MMs) have been proposed for the simulation of glucose -insulin metabolism for T1DM patients taking into consideration previous glucose measurements, type, and dose of insulin intake, and food intake, while recently, artificial neural networks (NN) have been proposed for simulation of glucose—insulin metabolism.

The aim of this article is to describe how NN have been applied for the simulation of glucose—insulin metabolism, and to present two NN based personalized models for children with T1DM. The models, which are able to make short-term glucose predictions, are based on the combined use of MMs and NNs. The models are comparatively assessed using data about glucose levels, insulin intake, and diet during previous time periods, from four children with T1DM.

BACKGROUND

Many computer-based simulations of glucose-insulin metabolism systems have been proposed in order to predict short-term glucose levels, and avoid hypoglycaemic and hyperglycaemic events. Usually, these systems, using among others glucose data from finger-stick glucose meters, are based on Compartmental Models (CMs). CMs represent a type of MMs commonly used for modeling complex dynamics systems, such as the physiological systems (Brown & Rothery, 1993; Cobelli & Foster, 1998). CMs have been used in order to estimate the way in which changes in insulin dosage affect blood glucose levels (Puckett & Lightfoot, 1995). Furthermore, a pharmacokinetic model has been developed by Berger and Rodbard (1989) for the simulation of plasma insulin and glucose dynamics after subcutaneous injection of insulin for the prediction of the expected time course of plasma glucose in response to a change in the insulin dose, timing or regime. This model has been improved by Lehmann and Deutsch (1992), by taking into consideration the absorption, in quantitative terms, of carbohydrates contained in the food. The widespread use of the resulted physiological model, known as AIDA, though being one of the most popular physiological models of glucose-insulin interaction for T1DM patients, it has been limited, due to uncertainties associated with the estimation of blood glucose profiles. These uncertainties have been addressed by Andreassen, Benn, Hovorka, Olesen, and Carson (1994) using causal probabilistic networks, for both the hourly prediction of blood glucose levels and the insulin dose adjustment. The obtained results have shown that this approach is able to identify and predict nocturnal hypoglucaemia (Cavan, Hovorka, Hejlesen, Andreassen, & Sönksen, 1996).

The acceptance of the aforementioned models was limited, because these systems take into account only a confined number of the factors associated with glucose metabolism, and they are not easily individualized to accurately simulate metabolic processes for a specific T1DM patient. In order to overcome the aforementioned difficulties, the use of NNs for the simulation of glucose-insulin metabolism has been proposed. NNs are a powerful tool in handling complex, nonlinear problems, due to their ability to be trained in order to handle unknown information hidden in the data (Haykin, 1999). Specifically, the information about blood glucose levels, insulin intake, and observed hypoglycemia symptoms have been used as input to a system consisting of two NN models in a chained scheme (Mougiakakou & Nikita, 2000; Mougiakakou, Nikita, Protonotarios, & Matsopoulos, 1998). The output of the first NN provides estimation for the insulin regime, while the output of the second NN estimate the appropriate insulin doses. Furthermore, inputs related to the insulin type and time of injection, carbohydrate and time of meal, exercise, blood glucose measurements, and special events (stress, pregnancy, and so on) from a T1DM patient have been fed to a NN resulting in a satisfactory accuracy in the prediction of blood glucose levels (Sandham, Hamilton, Japp, & Patterson, 1998). A comparative study of different NN approaches for the prediction of short-term blood glucose levels for T1DM patients has shown that the use of NNs is promising for the simulation of glucose metabolism (Tresp, Briegel, & Moody, 1999). Moreover, a hybrid system combining the principal component method with a NN has been proposed for the prediction of blood glucose levels in diabetic patients (Liszka-Hackzell, 1999). Recently, the use of a NN trained by the Levenberg-Marquardt (LM) algorithm has been presented, in order to predict the blood glucose concentration in T1DM patients (Zitar, 2003), with promising results. Finally, a hybrid model based on the combined use of MMs and NNs has been proposed for the short-term prediction of glucose levels in T1DM patients based on measured blood glucose levels, insulin intake, and description of food intake, along with the corresponding time (Mougiakakou, Prountzou, & Nikita, 2005).

MODELING OF GLUCOSE: INSULIN METABOLISM FOR CHILDREN WITH TYPE 1 DIABETES MELLITUS

In this article, two glucose-insulin metabolism models for children with T1DM are presented. The models, which are personalized and able to make short-term glucose predictions, are based on the combined use of CMs and NNs. The outline of the proposed models is shown in Figure 1. The models consist of a mathematical model (MM) module and a NN module. The MM module incorporates five CMs. The first four estimate plasma insulin concentration, due to quick acting (QA), short acting (SA), intermediate acting (IA), and long acting (LA) insulin intake, while the fifth the glucose absorption from the gut in response to carbohydrate intake. The outputs of the MM module along with the most recent glucose measurement are passed to the NN module, which provides prediction of short-term glucose levels.

Data Collection

Data from four children with T1DM have been used for the development and testing of the personalized glucose-insulin metabolism models. The characteristics of the patients are summarized in Table 1. Glucose data are collected using a CGMS (Continuous Glucose Monitoring Sensor, Medtronic MiniMed), which allow the measurement of interstitial glucose every five (5) minutes. The statistic of the glucose data is also given in Table I. The patients were monitored for a period between three to five days. For this period, the children recorded the time of insulin injection, the insulin type (QA, SA, IA, and LA) and dose, and the amount of carbohydrate ingested.

Mathematical Model Module

As aforementioned, the MM module consists of five CMs.

1. **Compartmental Models for Insulin Kinetics (CM-I, …, CM-V):** After the injection of D (U) units of insulin, the change in plasma insulin concentration $I(mU/lt)$ is given (Lehmann & Deutsch, 1992) as:

$$\frac{dI}{dt} = \frac{st^s (T_{50})^s D}{V_i[(T_{50})^s + t^s]t} - k_e I \qquad (1)$$

$$T_{50} = aD + b \qquad (2)$$

where $k_e = 5.4\ lt/h$ is the first-order rate constant of insulin elimination, $V_i = 9.94\ lt$ is the volume of insulin distribution, and T_{50} is the half- time insulin dose absorption. The constant parameters a, b, and s characterize insulin absorption patterns of the different insulin types entered in the model,

Figure 1. General architecture of the proposed artificial intelligent based glucose – insulin metabolism models

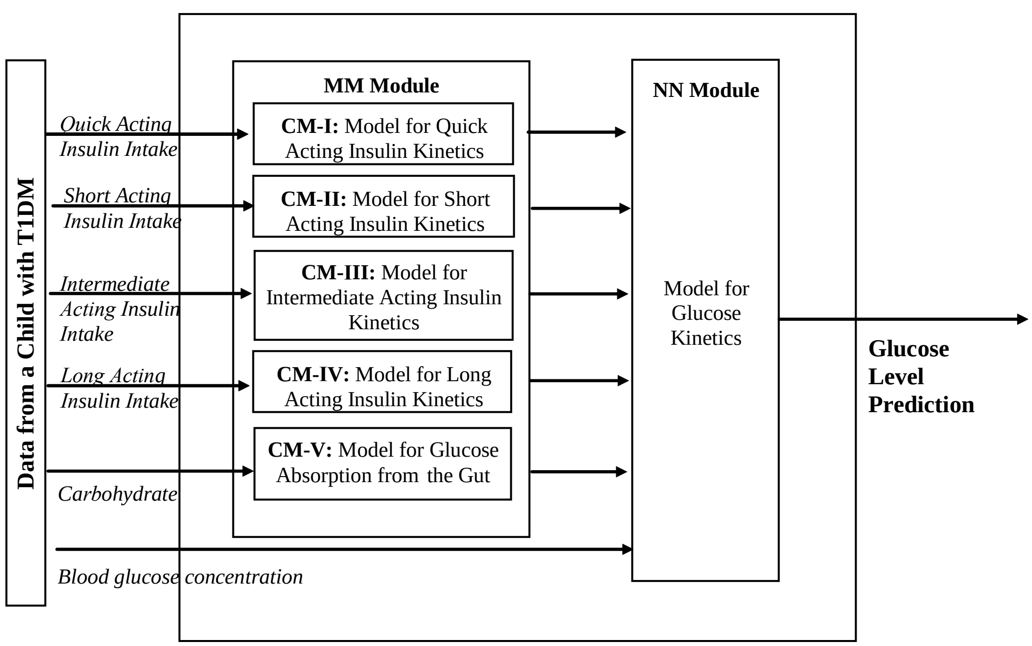


Table 1. Characteristics of the children with type 1 diabetes and statistics of the available glucose data (meas: measurements, SD: standard deviation)

Patient #	*Age (years)*	*Sex*	*Diabetes Duration (years)*	*BMI (kg/m^2)*	*Hba1c*	*# meas*	*Glucose levels (mg/dl)*		
							Min	*Max*	*Mean (SD)*
Patient 1	15	F	06	23.3	6.7	1311	44	400	157.07 (71.88)
Patient 2	15	F	13	29.5	7.5	1392	40	395	196.01 (94.53)
Patient 3	13	M	11	18.6	8.5	863	40	400	233.75 (96.61)
Patient 4	22	M	02	18.6	6.8	1069	57	228	123.03 (39.53)

and are given in (Lehmann, 1992). The differential equation (1) is solved using the fourth order Runge-Kutta method with a five-minute step.

2. **Compartmental model for glucose absorption from the Gut (CM-V):** After food intake, the glucose amount in the gut G_{gut} can be estimated by the equation:

$$\frac{dG_{gut}}{dt} = G_{empt} - k_{gabs} G_{gut} \qquad (3)$$

where $k_{gabs} = 1h^{-1}$ is the rate constant of glucose absorption from the gut into systemic circulation, and G_{empt} (mg / h) is the gastric emptying function given by a trapezoidal or triangular function, according to (Lehmann & Deutsch, 1996). Then, the glucose input into the blood from the gut wall is given by:

$$G_{in} = k_{gabs} \cdot G_{gut} \qquad (4)$$

Neural Network Module

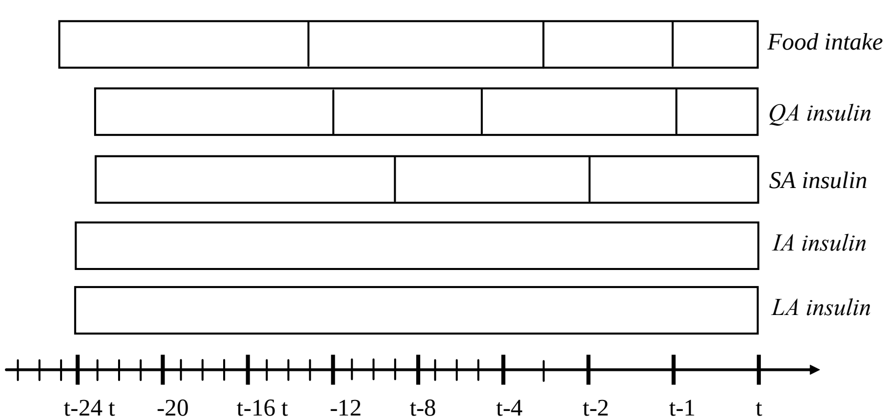
In order to predict the glucose level G (mg / dl) at a time instant t, the most recently measured blood glucose concentration along with parameters reflecting the effects of taken insulin and food are applied to the NN module. For comparative reasons, two different NN architectures have been developed and tested: a Feed-Forward NN (FFNN), and a Recurrent NN (RNN).

1. **Feed forward neural network:** The output of the FFNN is the glucose level G (mg / dl) prediction at time t, given by:

$$G(t) = \mathrm{FFNN}(G_{\mathrm{meas}}, \mathbf{G}_{\mathrm{in}}, \mathbf{I}_1, \mathbf{I}_2, \mathbf{I}_3, \mathbf{I}_4) \qquad (5)$$

where G_{meas} corresponds to the most recent glucose measurements, $\mathbf{G}_{\mathrm{in}} = \{G_{\mathrm{in},t-t_1}, G_{\mathrm{in},t-t_2}, G_{\mathrm{in},t-t_3}, G_{\mathrm{in},t-t_4}\}$ is a vector describing the effect of food intake in

Figure 2. Time windows (h) used for the description of insulin and food intake effects associated with blood glucose measurement at time t



the time-windows $t - t_i$ with i = 1, 2, 3, 4, $\mathbf{I}_1 = \{I_{1,t-t'_1}, I_{1,t-t'_2}, I_{1,t-t'_3}, I_{1,t-t'_4}\}$, is a vector describing the effect of taken QA insulin in the time-windows $t - t'_i$, $\mathbf{I}_2 = \{I_{2,t-t''_1}, I_{2,t-t''_2}, I_{2,t-t''_3}\}$, is a vector describing the effect of taken SA insulin in the time-windows $t - t''_i$ with i = 1, 2, 3, $\mathbf{I}_3 = I_{3,t-t'''_1}$ is a vector describing the effect of taken IA insulin in the time-window $t - t''_i$, and $\mathbf{I}_4 = I_{4,t-t''''_1}$ is a vector describing the effect of taken LA insulin in the time-window $t - t''''$. Each element of the above vectors is estimated as the sum of the prediction in the corresponding time windows. The used time-windows are indicated in Figure 2, and have been selected according to the expertise of physicians (Tresp et al., 1999). If in a time-window more than one event of the same type occurs, then the combined effect of the events is taken into account.

In Equation 5, the symbol FFNN denotes the nonlinear transformation of the data according to the used NN architecture. A fully connected feed-forward NN, trained by the batched back-propagation algorithm with adaptive learning rate and momentum (Haykin, 1999), has been used. The NN consists of three layers: input, hidden, and output layer. The input layer is formed by a number of neurons equal to the number of parameters used for the description of previous blood glucose measurements, insulin, and food intake. The hidden layer consists of a variable number of neurons ranging from three to 25. The output layer consists of one neuron, providing prediction of the subsequent blood glucose level. The tan-sigmoid function has been used for both hidden and output layers. In order to define the optimal number of neurons, the momentum term, and the initial value of learning rate, a trial-and-error process has been followed until no further improvement in glucose prediction could be obtained.

2. **Recurrent neural network:** In the case of RNN, the outputs of the MM module provide exogenous input to the RNN, which is a second-order RNN with one state variable. The output of the RNN is the glucose level G (gm / dl) prediction at time t, given by:

$$G(t) = G(t-1) + \mathrm{RNN}(G(t-1), G(t-2), G_{in}, I_1, I_2, I_3, I_4) \tag{6}$$

where $G(t-1)$, $G(t-2)$ are the current and previous blood glucose predictions, respectively, G_{in} is the glucose input into the blood from the gut, I_1 is the plasma insulin concentration due to QA insulin intake, and I_2 is the plasma insulin concentration due to SA insulin intake, I_3 is the plasma insulin concentration due to IA insulin intake, and I_4 is the plasma insulin concentration due to LA insulin intake.

The RNN is a fully connected NN trained with the online Real Time Recurrent Learning (RTRL) algorithm (Haykin, 1999). The used training algorithm has the ability to update online the RNN weights. As long as the RNN is provided with new inputs (external and internal), it adapts the weights accordingly, trying to simulate the dynamic system at real time. For comparative reasons, two strategies have been applied: the Free-Run (FR), and the Teacher-Forcing (TF). In the case of FR (online RTRL–FR), the RNN ignores the available glucose measurement, while in the case of TF (online RTRL–TF) strategy, the RNN replace the actual output, during training, with the corresponding available glucose measurement.

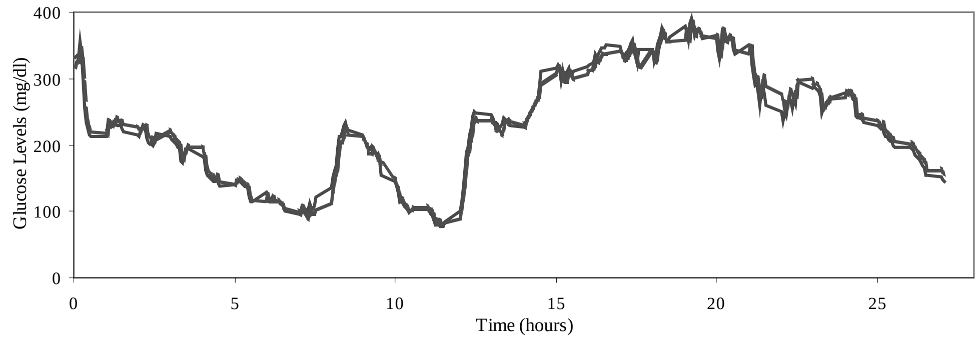
The RNN consists of three layers: the input layer formed by five external neurons, the hidden layer with a number of neurons ranging from five to 15, and the output layer with one neuron, providing prediction of the change of blood glucose level at next time step. The tan-sigmoid and linear activation functions have been used for the hidden and the output layer, respectively. The optimal configuration of RNN is defined following the same procedure as for FFNN.

Results and Discussion

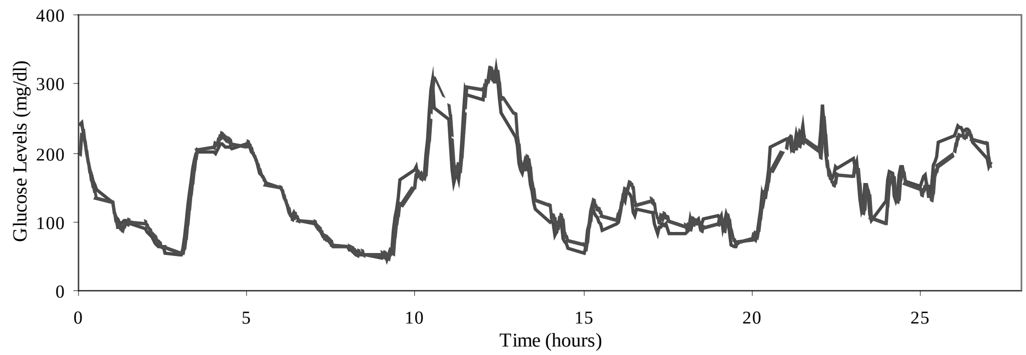
Aiming at the minimization of overtraining risk, and at the maximization of the models ability to predict glucose levels, the recorded days have been divided into two disjoint datasets: training, and testing sets. For each patient, the data of the first days (minimum three days) using CGMS have formed the training set, and the data of the last day using CGMS the testing set.

In order to assess the performance of the developed models, the Root Mean Square Error (RMSE) and the correlation coefficient (cc) have been calculated for both data sets. In Table 2, the RMSE, and the cc are given for each model.

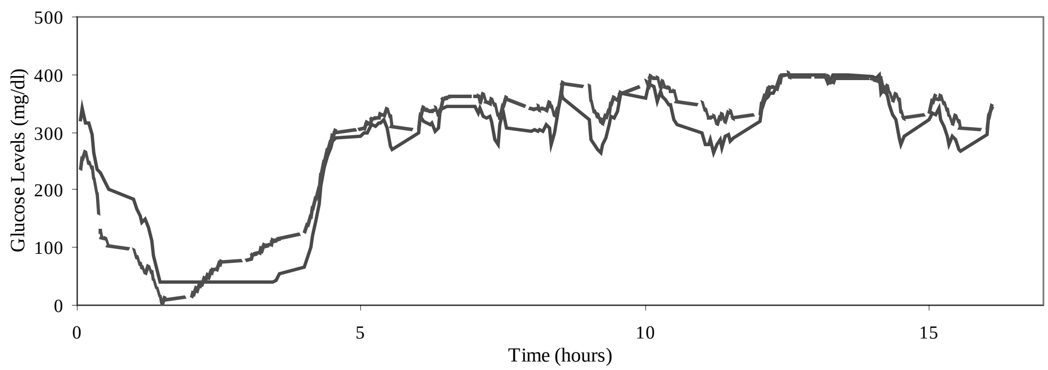
Figure 3. Comparison between estimated by the model using the online RTRL – TF trained RNN (-) and the measured (-) glucose levels for the testing set, in case of (a) patient 1, (b) patient 2, (c) patient 3, and (d) patient 4



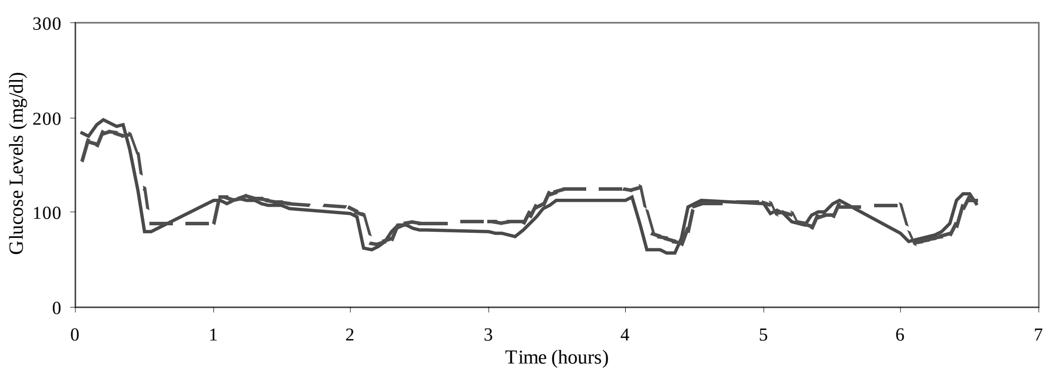
(a)



(b)



(c)



(d)

From Table 2, it can be shown that the results obtained from the FFNN and the RNN trained with the online RTRL-TF are superior to those obtained by the online RTRL-FR trained RNN for all diabetes patients. Furthermore, it can be observed that the performance of FFNN is slightly better, as compared to the online RTRL-TF trained RNN. Despite this, the online RTRL-TF trained RNN is preferable, due to its ability to adapt the weights whenever a new input is applied.

Finally, in Figure 3, the measured and the estimated blood glucose levels vs. time are presented for the testing set of all patients, using online RTRL-TF trained RNN. From the above results, it is obvious that further studies need to be performed, including a greater number of T1DM patients in order to improve the significance of the results.

FUTURE TRENDS

Future direction of the current research includes the development of an automatic advisor, which can make recommendations about the appropriate (i) time and the doses of insulin for patient using MDI, and (ii) insulin pumps settings for patient using CSII is in progress. The system will be part of an integrated platform for the monitoring and management of T1DM patients. The platform will be based on the combined use of insulin pumps, blood glucose measurement devices, internet technology, Wireless Personal Area Networks (WPAN), mobile communication networks, and tools for the intelligent processing of diabetes information.

The ultimate goal is the development of an autonomous closed-loop system able to mimic accurately the glucose-insulin metabolism, and maintain the glucose levels into physiological range.

Table 2. Root mean square errors (RMSE), along with correlation coefficients (cc) between measured and estimated by the models glucose levels for the testing set

Patient #	*Model*	*RMSE*	*cc*
Patient 1	*FFNN*	27.82	0.94
	RNN/FR	39.32	0.80
	RNN/TR	15.13	0.97
Patient 2	*FFNN*	7.19	0.99
	RNN/FR	38.11	0.91
	RNN/TR	11.58	0.99
Patient 3	*FFNN*	41.34	0.98
	RNN/FR	71.69	0.79
	RNN/TR	55.38	0.90
Patient 4	*FFNN*	12.31	0.92
	RNN/FR	33.47	0.62
	RNN/TR	14.25	0.91

CONCLUSION

Advancements in information technology and software engineering permit the design and development of computational models able to simulate accurately complex, nonlinear biological, and physiological processes. In this article, the application of NNs for the design and development of glucose–insulin metabolism models for T1DM patients has been presented. Furthermore, two glucose-insulin metabolism models for children with T1DM have been described. The models, which are based on the combined use of CMs and NNs, provide individualized, accurate, short-term predictions of glucose levels.

ACKNOWLEDGMENT

This work is being supported in part by the General Secretariat of Research and Technology/Ministry of Development (Greece) under the R&D project "SMART-DIAB: Insulin Infusion Intelligent System."

REFERENCES

Andreassen, S., Benn, J.J, Hovorka, R., Olesen, K.G., & Carson, E.R. (1994). A probabilistic approach to glucose prediction and insulin dose adjustment: Description of metabolic model and pilot evaluation study. *Computer Methods and Programs in Biomedicine, 41*, 153–165.

Berger, M.,& Rodbard, D. (1989). Computer simulation of plasma insulin and glucose dynamics after subcutaneous insulin injection. *Diabetes Care, 12*(10), 725–736.

Bequette, B.W. (2005). A critical assessment of algorithms and challenges in the development of a closed-loop artificial pancreas. *Diabetes Technology and Therapeutics, 7*(1), 28–47.

Brown, D., & Rothery, P. (1993). *Models in biology: Mathematics, statistics and computing*. John Wiley & Sons.

Cavan, D.A., Hovorka, R., Hejlesen, O.K., Andreassen, S., & Sönksen, P.H. (1996). Use of the DIAS model to predict unrecognised hypoglycaemia in patients with insulin-dependent diabetes. *Computer Methods and Programs in Biomedicine, 50*, 241–246.

Cobelli, C., & Foster, D.M. (1998). Compartmental models: Theory and practice using the SAAM II software system. *Advances in Experimental Medicine and Biology, 445*, 79–101.

DCCT Research Group. (1993). The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *New England Journal in Medicine, 329*(14), 977–986.

Haykin, S. (1999). *Neural networks: A comprehensive foundation*. Prentice Hall.

Lehmann, E.D., & Deutsch, T. (1992). A physiological model of glucose-insulin interaction in type 1 diabetes mellitus. *Journal of Biomedical Engineering, 14*, 235–241.

Lehmann, E.D., & Deutsch, T. (1996). Computer Assisted Diabetes Care: A 6-year Retrospective. Computer Methods and Programs in Biomedicine, 50, 209–230.

Liszka-Hackzell, J.J. (1999). Prediction of blood glucose levels in diabetic patients using a hybrid ai technique. *Computers and Biomedical Research, 32,* 132–144.

Mougiakakou, S.G., & Nikita, K.S. (2000). A neural network approach for insulin regime and dose adjustment in type 1 diabetes. *Diabetes Technology & Therapeutics, 2*(3), 381–389.

Mougiakakou, S.G., Nikita, K.S., Protonotarios, M., & Matsopoulos, G. (1998). Neural network based decision support system for outpatients diabetes management. *Diabetes, Nutrition & Metabolism, 11*(1), 74.

Mougiakakou, S.G., Prountzou, K., & Nikita, K.S. (2005). A real time simulation model of glucose-insulin metabolism for type 1 diabetes patients. *Engineering in Medicine and Biology Conference 2005* (pp. 298–301). Shanghai, China.

Puckett, W.R., & Lightfoot, E.N. (1995). A model for multiple subcutaneous insulin injections developed from individual diabetic patient data. *American Journal of Physiology, 269,* 1115–1124.

Sandham, W.A., Hamilton, D.J., Japp, A., & Patterson, K. (1998). Neural network and neuro-fuzzy systems for improving diabetes therapy. In *Proceedings of the 20th Annual International Conference IEEE/EMBS* (pp. 1438–1441). Hong-Kong, China.

Tresp, V., Briegel, T., & Moody, J. (1999). Neural network models for the blood glucose metabolism of a diabetic. *IEEE Transactions on Neural Networks, 10*(5), 1204–1213.

UKPDS Group. (1998). Intensive blood glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS33). *Lancet, 352*, 837–853.

Zitar, R.A. (2003). Towards neural network model for insulin/glucose in diabetics. *International Journal of Computing and Information Sciences, 1*(1), 25–32.

KEY TERMS

Artificial Neural Networks: Neural networks that include artificial intelligence.

Compartmental Models: Models that have distinct sections.

Continuous Glucose Monitoring Systems: Systems that monitor glucose levels at all times.

Personalized Models: Models that are individualised.

Type 1 Diabetes Mellitus: A condition characterized by disordered metabolism and inappropriately high blood sugar resulting from either low levels of the hormone insulin, or from abnormal resistance to insulin's effects, coupled with inadequate levels of insulin secretion to compensate.

Integrating Medication and Health Monitoring Systems

Burges Karkaria
Intel Corporation, USA

Vincent Zimmer
Intel Corporation, USA

INTRODUCTION

Current health care providers (medical practitioners, insurance companies, pharmacies) have disjointed proprietary systems. This makes data sharing and collaboration very difficult. Furthermore, consumers (patients) have no abilities to interact directly with these systems. This chapter discusses an ecosystem for patient medication and health monitoring called the patient medication and health monitoring system (PMHMS). PMHMS proposes global and ongoing collaboration of data between the patient, various medical practitioners, and pharmacies. Existence of this ecosystem helps to enable some of the following scenarios:

- Medical practitioners can view the patient's current medication (from other medical practitioners as well) to ensure that the current prescription will not conflict with other medications.
- The medical practitioner's office can monitor if chronic patient is skipping medication (e.g., no refills ordered). The medical practitioner's office can also monitor patient's stats (patient uses PC to upload, for instance, blood sugar levels, blood pressure, and so forth).
- Pharmacies will be able to automatically send refills to patients.
- Patients can get a quick off-site consultation from the medical practitioner's office (using video conferencing to show some visual symptoms as needed by the medical practitioner)
- Patients get automated dosage reminders.

BACKGROUND

A consumer facing commerce model generally consists of the consumer, main and ancillary services, and a financial broker or approver. As an example, we look at some sample models. Figure 1 shows the constituents and associated transactions of a consumer electronics store, while Figure 2 shows the same for a travel agency. Both of these depict seamless end-consumer experience.

In comparison, Figure 3 shows an example model of today's health care systems. While this example is more relevant to certain geographies like North America and Europe, it does highlight the lack of integration in the health care system constituents. The PMHMS system is a sample implementation aiming to rectify some of these issues.

PATIENT MEDICATION AND HEALTH MONITORING SYSTEM (PMHMS)

Overview

PMHMS is an ecosystem for patient medication and health monitoring. There is global and ongoing collaboration of data between the patient, various medical practitioners, and pharmacies. Existence of this ecosystem helps to enable some of the following scenarios (for a complete list of PMHMS possibilities, see Use Cases in Figure 4):

- Medical practitioners can view the patient's current medication (from other medical practitioners as well) to ensure that the current prescription will not conflict with other medications.
- The medical practitioner's office can monitor if chronic patient is skipping medication (e.g., no refills ordered). The medical practitioner's office can also monitor patient's stats (patient uses PC to upload, for instance, blood sugar levels, blood pressure, and so forth).

Figure 1. Consumer store economic model constituents

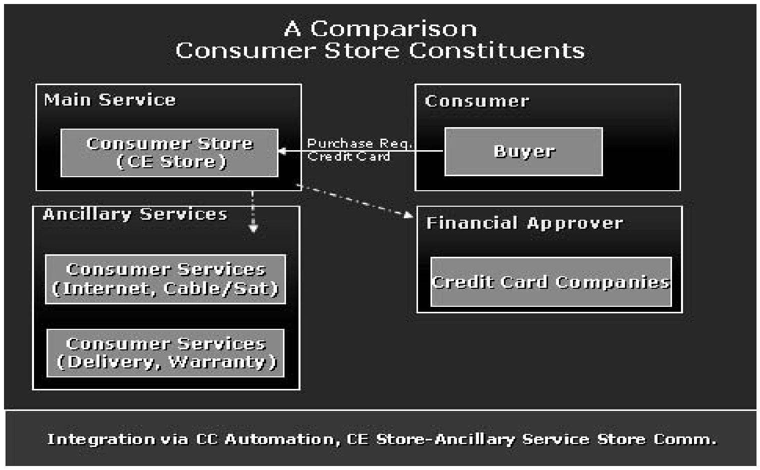
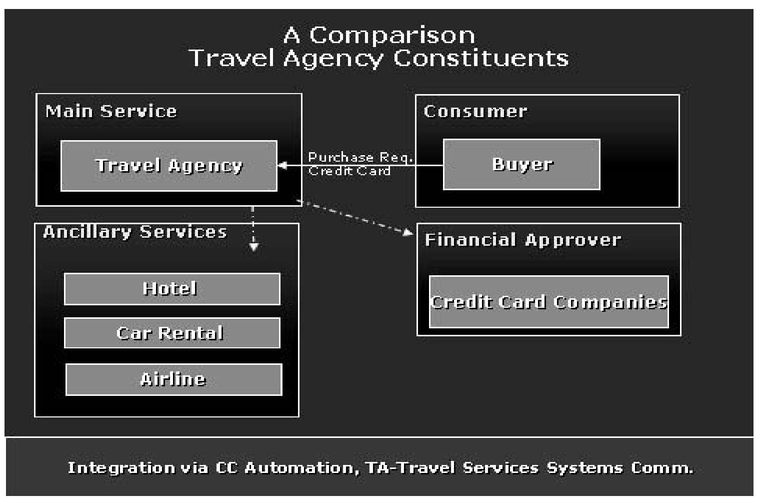


Figure 2. Travel agency economic model constituents



- Pharmacies will be able to automatically send refills to patients.
- Patients can get a quick off-site consultation from the medical practitioner's office (using video conferencing to show some visual symptoms as needed by the medical practitioner)
- Patients get automated dosage reminders

PMHMS Subsystems

PMHMS consists of the subsystems as shown in Figure 5.

Central Patient Repository (CPR)

CPR is a centrally available database of the medication history of the patient. (This repository also can be upgraded later to include a patient's medical records (e.g., previous diagnosis, X-ray scans, blood reports, etc.). CPR is run as a commercial service. Multiple vendors may be offering CPR services, and the insurance provider may choose to use its vendor of choice based on factors such as cost, features, security, and so forth. The system is built on an established set of open standards (e.g., HL7 [Std1]).

Figure 3. Health care system constituents

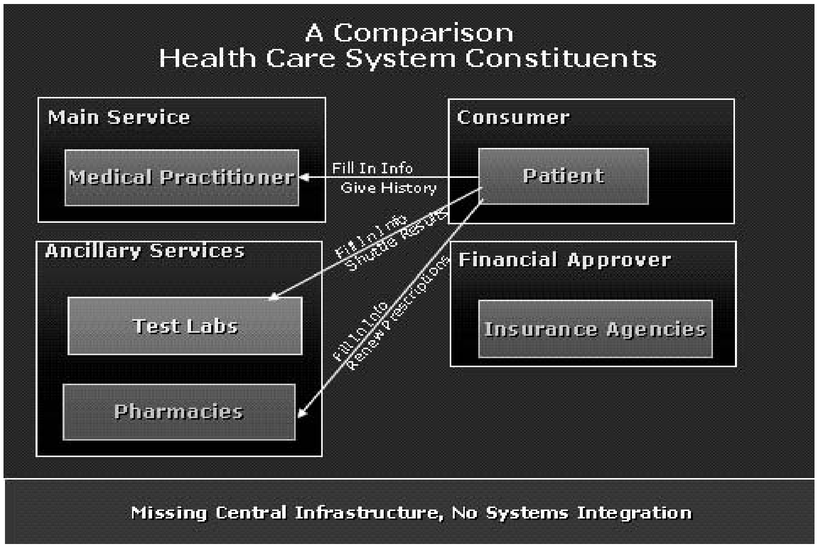
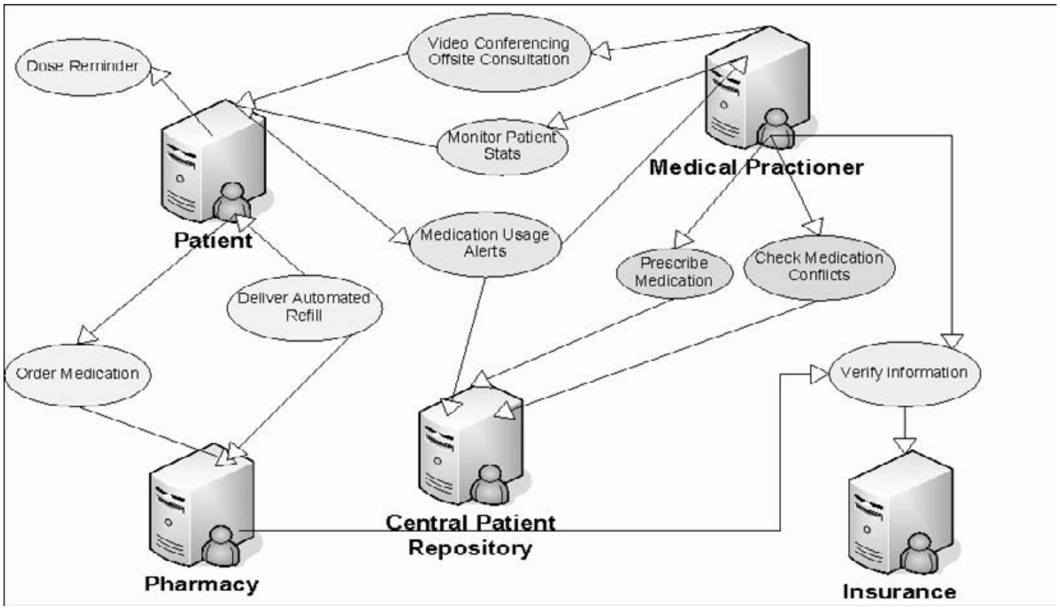


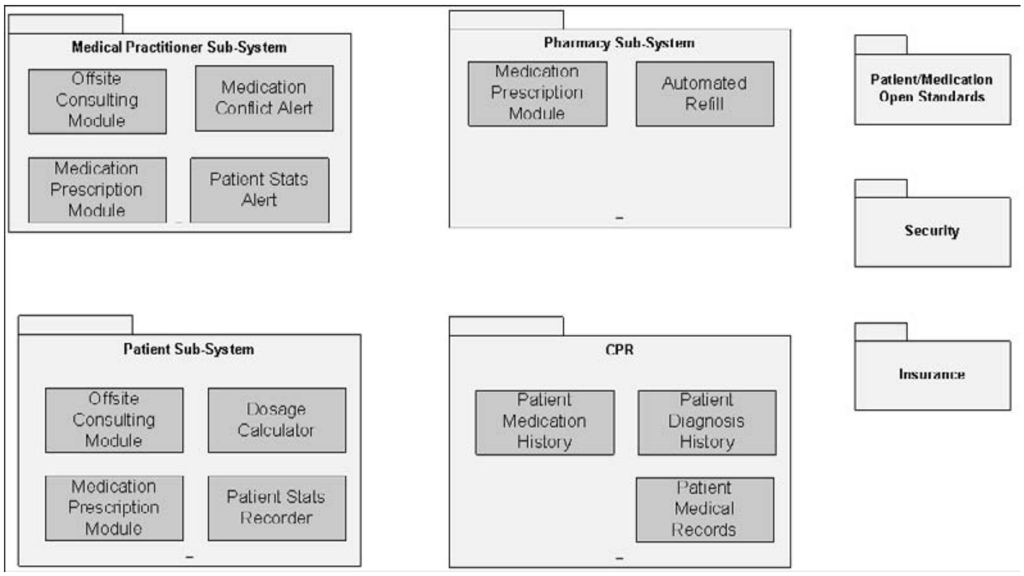
Figure 4. PMHMS possibilities/use cases



Medical Practitioner's Subsystem

The Medical Practitioner's Subsystem is a local app installed in the medical practitioner's office. The medical practitioner can use this subsystem to view the patient's medication history (using CPR). The medical practitioner also uses this to log prescriptions into the CPR. This subsystem also has the ability to detect potentially conflicting drugs. If the Medical Practitioner is prescribing some medication that may conflict with another prescribed drug, the system will alert the medical practitioner, and the medical practitioner has the ability to change the prescription.

Figure 5. PMHMS subsystems



Pharmacy Subsystem

The Pharmacy Subsystem takes medication orders, gets refills sent out automatically as applicable, and so forth.

Insurance Subsystem

The Insurance Subsystem lets medical practitioners and pharmacies verify patient information

Patient Subsystem

The Patient Subsystem is a local app installed on the patient's local PC. Using the Patient Subsystem, the consumer can do some of the following:

- **Ordering medication:** PMHMS allows the user to order prescription drugs from pharmacies. The user can have a preferred pharmacist (in which case new prescriptions automatically get delivered). It automatically notifies the pharmacy for refills when one is needed.
- **Medication intake and health monitoring:** This system also prompts the user about what dosage needs to be taken next. The system also allows users to log into their statistics such as blood sugar levels or blood pressure (using approved devices that connect to the PC). If the statistics show an undesirable spike, then an alert goes out to the relevant medical practitioner, and the medical practitioner's office can follow up to ensure the patient's well-being.

Finally, the patient can have small off-site consultation sessions directly with the medical practitioner's office by starting video conferencing and VOIP sessions.

Use of Standards

The reason various subsystems can communicate with each other is due to the use of standards. The standards are applied via the following:

- **Data standards:** Medication, diseases, patient, and so forth are well defined (we could use existing standards such as HL7 (health Level 7) [Std1] or DICOM.
- **Service standards:** The services that need to be exposed by the subsystems need to be standardized as well. There could be internal differentiation in the implementation of the service. For example, all CPR service providers should offer a service such as GetPatientCurrentMedication (Patient, MedicationColl).
- **Privacy standards:** Going to a more automated electronic system of medical records handling

raises concerns about patient privacy. The Health Insurance Portability and Accountability Act of 1996 [Std3] was intended to provide guidelines to avoid fraud, waste, and other information abuses. This is key to assuage customer concerns for this medical information automation.

FUTURE TRENDS

The evolution of standards is going to continue due to focused government and industry initiatives. If certain health care providers start providing some of the automation depicted previously, it will be a huge differentiating factor enhancing their basic service offerings. This in turn would fuel a competitive race to provide equivalent customer service via integration. The end result would be a highly automated model facilitating myriad advances in patient health care via specialized appliances and services.

CONCLUSION

Health care systems today are plagued by lack of integration. This causes huge end-user disadvantages such as redundant information filling, increase in stress, and so forth. It can also be the cause of critical and costly mistakes such as medication conflicts as a result of incorrect medication history disclosure. An integrated ecosystem built with an end-consumer centric focus can eliminate these issues. The enabling of Health Care Systems Integration can be caused by the use of existing and evolving standards.

REFERENCES

http://www.hl7.org

http://www.ihe.net

http://www.hipaa.org/

http://medical.nema.org/dicom/

KEY TERMS

CPR: Central Patient Repository.

DICOM: Digital Imaging and Communications is Medicine.

Ecosystem: An existing infrastructure of automated systems.

HIPAA: Health Insurance Portability and Accountability Act.

HL7: Health Level 7.

IHE: Integrating Health Care Enterprise.

Medical Practitioner: Could be a general physician, chiropractor, dentist, specialist, surgeon, etc.

PMHMS: Patient Medication and Health Monitoring System.

Integration of Clinical and Genomic Data for Decision Support in Cancer

Yorgos Goletsis
University of Ioannina, Greece

Themis P. Exarchos
University of Ioannina, Greece

Nikolaos Giannakeas
University of Ioannina, Greece

Markos G. Tsipouras
University of Ioannina, Greece

Dimitrios I. Fotiadis
University of Ioannina, Greece
Michaelideion Cardiology Center, Greece & Biomedical Research Institute, Greece

INTRODUCTION

Computer aided medical diagnosis is one of the most important research fields in biomedical engineering. Most of the efforts made focus on diagnosis based on clinical features. The latest breakthroughs of the technology in the biomolecular sciences are a direct cause of the explosive growth of biological data available to the scientific community. New technologies allow for high volume affordable production and collection of information on biological sequences, gene expression levels and proteins structure on almost every aspect of the molecular architecture of living organisms. For this reason, bioinformatics is asked to provide tools for biological information processing, representing today's key in understanding the molecular basis of physiological and pathological genotypes. The exploitation of bioinformatics for medical diagnosis appears as an emerging field for the integration of clinical and genomic features, maximizing the information regarding the patient's health status and the quality of the computer aided diagnosis.

Cancer is one of the prominent domains where this integration is expected to bring significant achievements. As genetic features play a significant role in the metabolism and the function of the cells, the integration of genetic information (proteomics-genomics) to cancer-related decision support is now perceived by many not as a future trend but rather as a demanding need. The usual patient management in cancer treatment involves several, usually iterative, steps consisting of diagnosis, staging, treatment selection, and prognosis. As the patient is usually asked to perform new examinations, diagnosis and staging status can change over time. On the other hand, treatment selection and prognosis depend on the available findings, response to previous treatment plan and, of course, clinical guidelines. The integration of these evolving and changing data into clinical decision is a hard task which makes fully personalised treatment plan almost impossible. The use of clinical decision support systems (CDSSs) can assist in the processing of the available information and provide accurate staging, personalised treatment selection, and prognosis. The development of electronic patient records and of technologies that produce and collect biological information have led to a plethora of data characterizing a specific patient. Although this might seem beneficial, it can lead to confusion and weakness concerning the data management. The integration of the patient data (quantitative) that are hard to be processed by a human decision maker (the clinician) further imposes the use of CDSSs in personalized medical care (Louie, Mork, Martin-Sanchez, Halevy, & Tarczy-Hornoch, 2007). The future vision—but current need—will not include generic treatment plans according to some naive reasoning, but totally personalised treatment based on the clinicogenomic profile of the patient.

In this article, we address decision support for cancer by exploiting clinical data and identifying mutations on tumour suppressor genes. The goal is to perform data integration between medicine and molecular biology by developing a framework where clinical and genomic features are appropriately combined in order to handle cancer diseases. The constitution of such a decision support system is based on (a) cancer clinical data and (b) biological information that is derived from genomic sources. Through this integration, real time conclusions can be drawn for early diagnosis, staging and more effective cancer treatment.

BACKGROUND

Clinical Decision Support Systems are active knowledge systems which use two or more items of patient data to generate case-specific advice (Fotiadis, Goletsis, Likas, & Papadopoulos, 2006). CDSSs are used to enhance diagnostic efforts and include computer based programs that, based on information entered by the clinician, provide extensive differential diagnosis, staging (if possible), treatment, follow-up, and so forth. CDSSs consist of an inference engine that is used to associate the input variables with the target outcome. This inference engine can be developed based either on explicit medical knowledge, expressed in a set of rules (knowledge based systems) or on data driven techniques, such as machine learning (Mitchel, 2006) and data mining (intelligent systems) (Tan, Steinbach, & Kumar, 2005). CDSSs require the input of patient-specific clinical variables (medical data) and as a result provide patient specific recommendation.

Medical data are observations regarding a patient, including demographic details (i.e., age, sex), medical history (i.e., diabetes, obesity), laboratory examinations (e.g., creatinine, triglyceride), biomedical signals (ECG, EMG), medical images (i.e., MRI, CT), and so forth. Demographic details, medical history, and laboratory data are the most easily obtained and recorded and, therefore, most commonly included in electronic patient records. On the other hand, biomedical signals and medical images require more effort in order to be acquired in a digital format and must be processed for useful feature extraction. Apart from these, several types of genomic data can be generated from laboratory examinations, that is, gene DNA or protein sequences, gene expression data, microarray images, and so forth. Genomic data can also be used for medical diagnosis, disease prevention, and population genetics studies. Although medical data are sufficient for the diagnosis of several diseases, recent studies have demonstrated the high information value of genomic data, especially in specific types of diseases, such as cancer diseases.

The great amount and the complexity of the available genetic data complicates their analysis from conventional data analysis methods and requires higher order analysis methods such as data mining techniques. Lately, data mining has received much attention in bioinformatics and molecular biology (Cook, Lawrence, Su, Maglothin, & Jonyer, 2001). Data mining methods are usually applied in the analysis of data coming from DNA microarrays or mass spectrometry. Over the last few years, several scientific reports have shown the potential of data mining to infer clinically relevant models from molecular data and thus provide clinical decision support. The majority of papers published in the area of data mining for genomic medicine deals with the analysis of gene expression data coming from DNA microarrays (Jiang & Gruenwald, 2005; Shah & Kusiak, 2007; Walker et al., 2004) consisting of thousands of genes for each patient, with the aim to diagnose types of diseases and to obtain a prognosis which may lead to therapeutic decisions. Most of the research works are related to oncology (Louie et al., 2007), where there is a strong need for defining individualized therapeutic strategies. Another area where data mining has been applied is the analysis of genes or proteins, represented as sequences (Exarchos, Papaloukas, Lampros, & Fotiadis, 2006); sequential pattern mining techniques are suitable for analyzing these types of data (Zaki, 2000).

Several CDSSs for cancer have been proposed in the literature. Most of the approaches are based solely on clinical data and a few methods exist that provide cancer decision support using microarray gene expression data. The cancer CDSSs concern several different types of cancer and employ various techniques for their development. The majority of systems are still in a research level and only a few are being used in clinical practice. A CDSS which is already in clinical use is PAPNET (Boon & Kok, 2001) which deals with cervical cancer. PAPNET uses ANNs to extract abnormal cell appearances from vaginal smear slides and describe them in histological terms. Other CDSSs for cervical cancer concentrate on the evaluation of the benefits of the PAPNET system (Doornewaard, 1999; Nieminen, Hakama, Viikki, Tarkkanen, & Anttila, 2003). Colon cancer has also been studied using clinical data and fuzzy classification trees (Chiang, Shieh, Hsu, & Wong, 2005) or pattern analysis of gene expression levels

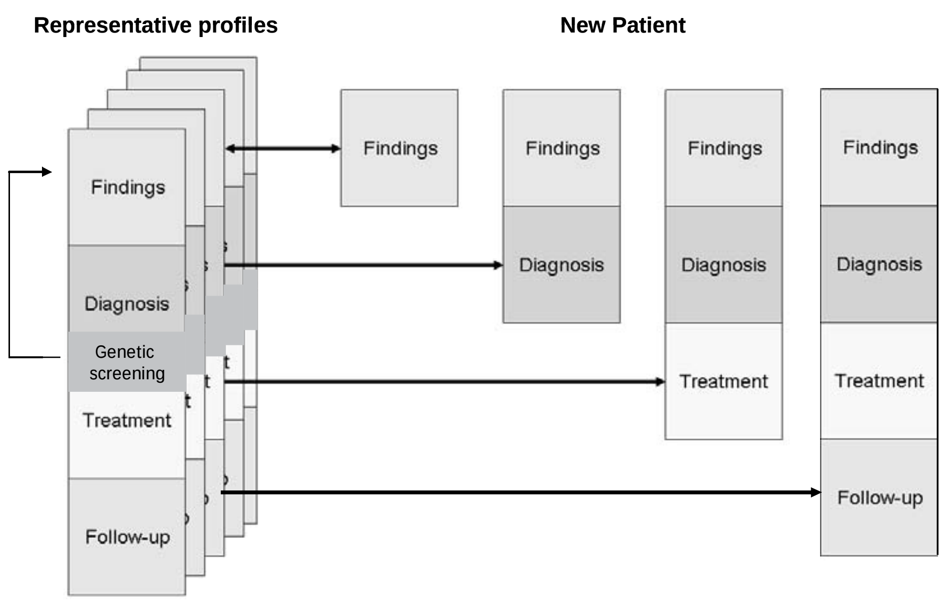
(Alon et al., 1999). A CDSS that combines imaging data with pathology data for colon cancer has also been proposed (Slaymaker, 2006). CDSSs proposed for prostate cancer, employ prostate specific antigen (PSA) serum marker, digital rectal examination, Gleason sum, age, and race (Remzi et al., 2003). Another approach for decision support in prostate cancer is based on gene expression profiles (Singh et al., 2002). Regarding bladder cancer, a CDSS has been developed based on proteomic data (Parekattil, Fisher, & Kogan, 2003). Concerning breast cancer, the potential of microarray data has been analysed (Van't Veer et al., 2002). Also, a recent CDSS has been developed that integrates data mining with clinical guidelines towards breast cancer decision support (Skevofilakas, Nikita, Templakesis, Birbas, Kaklamanos, & Bonatsos, 2005). It should be noted that all CDSSs mentioned above are just research attempts and only PAPNET is in clinical use.

CLINICAL DECISION SUPPORT USING CLINICOGENOMIC PROFILES

Methodology

Conventional approaches for CDSS focus on a single outcome regarding their domain of application. A different approach is to generate profiles associating the input data (e.g., findings) with several different types of outcomes. These profiles include clinical and genomic data along with specific diagnosis, treatment and follow-up recommendations. The idea of profile-based CDSS is based on the fact that patients sharing similar findings are most likely to share the same diagnosis and should have the same treatment and follow-up; the higher this similarity is, the more probable this hypothesis holds. The profiles are created from an initial dataset including several patient cases using a clustering method. Health records of diagnosed and (successfully or unsuccessfully) treated patients, with clear follow-up description, are used to create the profiles. These profiles constitute the core of the CDSSs; each new case that is inserted is related with one (or more) of these profiles. More specifically, an individual health record containing only findings (and maybe the diagnosis) is matched to the centroids. The matching centroids are examined in order to indicate potential diagnosis (the term diagnosis here refers mainly to the identification of cancer subtype). If the diagnosis is confirmed, genetic screening may be proposed to the subject and then, the clusters are further examined, in order to make a decision regarding the preferred treatment and follow-up. The above decision support idea is shown schematically in Figure 1.

Figure 1. Decision support based on profiles extraction. Unknown features of patient are derived by known features of similar cases.



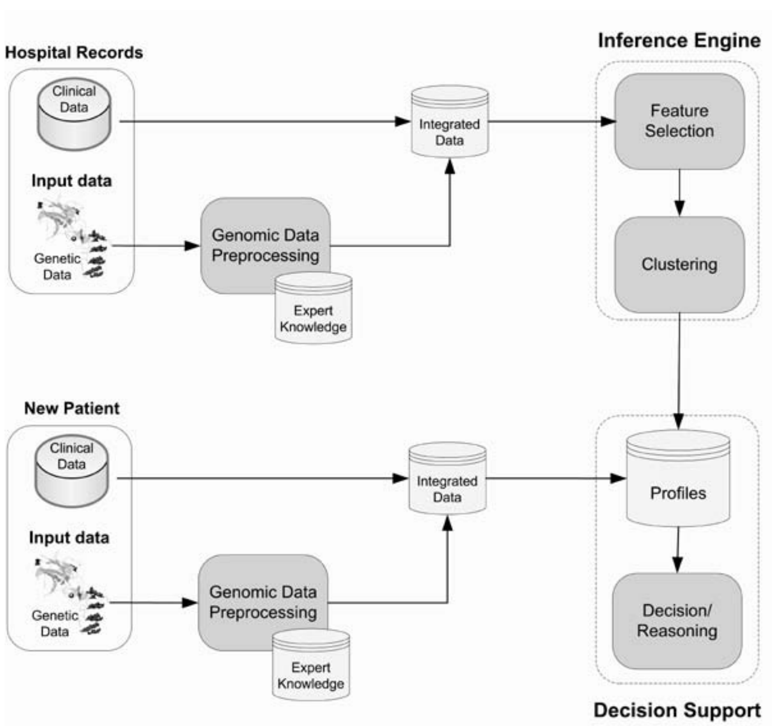
General Description of the System

Known approaches for the creation of CDSSs are based on the analysis of clinical data using machine learning techniques. This scheme can be expanded to include genomic information, as well. In order to extract a set of profiles, the integration of clinical and genomic data is first required. Then data analysis is realized in order to discover useful knowledge in the form of profiles. Several techniques and algorithms can be used for data analysis such as neural approaches, statistical analysis, data mining, clustering and others. Data analysis is a two stage procedure: (i) creation of an inference engine (training stage) and (ii) use of this engine for decision support. The type of analysis to be used greatly depends on the available information and the desired outcome. Clustering algorithms can be employed in order to extract patient clinico-genomic profiles. An initial set of records, including clinical and genomic data along with all diagnosis/treatment/follow-up information, must be available for the creation of the inference engine. The records are used for clustering and the centroids of the generated clusters constitute the profiles. These profiles are then used for decision support; new patients with similar clinical and genomic data are assigned to the same cluster, that is, they share the same profile. Thus, a probable diagnosis, treatment, and follow-up is selected. Both, the creation of the inference engine and the decision support procedure are presented in Figure 2.

Types of Data

Clinical data that are contained in electronic patient records (i.e., demographic details, medical history, and laboratory data) are usually presented in a simple and structured format, thus simplifying the analysis. On the other hand, genomic data are not structured and, therefore, appropriate preprocessing is needed in order to transform them into a more structured format. Three different kinds of biological data may be available to clinicians: (i) genomic data, often represented by a collection of single nucleotide polymorphisms (SNPs), DNA sequence variations that occur when a single nucleotide in the genome sequence is altered and since each individual has many SNPs, their sequence forms a unique DNA pattern for that person; (ii) gene expression data, which can be measured with DNA microarrays to obtain a snapshot of the activity of all genes in one tissue at a given time or with techniques

Figure 2. Representation of a general scheme for a CDSS, integrating clinical and genomic information



that rely on a polymerase chain reaction (PCR) and real-time PCR when the expression of only a few genes needs to be measured with better precision; (iii) protein expression data, which can include a complete set of protein profiles obtained with mass spectra technologies, or a few protein markers which can be measured with ad hoc essays.

Data Processing

Depending on the type of the available biological data, different preprocessing steps should be performed in order to derive structured biological information, while expert knowledge could favor the preprocessing steps. The processing stage is necessary in order to transform the genomic data into a more easy-to-analyse form, allowing their integration along with the clinical data into the data analysis stage. Also, the genomic data processing might take advantage of expert knowledge, that is, known genomic abnormalities. Finally, the integrated data (clinical and genomic) are analysed in order to discover useful knowledge that can be used for decision support purposes. This knowledge can be in the form of associations between clinical and genomic data, differential diagnosis, treatment, and so forth.

The initial dataset (clinical or genomic) is defined by the experts and includes all features that according to their opinion are highly related with the domain at hand (clinical disease). After acquiring the integrated data, a feature selection technique is applied in order to reduce the number of features and remove irrelevant or redundant ones. Finally, the reduced set of features is used by a clustering algorithm. K-means (MacQueen, 1967), fuzzy k-means (Bezdek, 1981), and expectation maximization (Dempster, Laird, & Rubin, 1977) are known approaches for clustering and can be involved for profile extraction. The profiles are the output of the clustering procedure (centroids). A deficiency of several clustering algorithms is that the number of centroids (profiles) must be predefined; this is not always feasible. Thus, in order to fully automate the profile extraction process, a meta-analysis technique is employed, which automatically calculates the optimal number of profiles.

Application to Colon Cancer

Colon cancer includes cancerous growths in the colon, rectum, and appendix. It is the third most common form of cancer and the second leading cause of death among cancers in the developed countries. There are many different factors involved in colon carcinogenesis. The association of these factors represents the base of the diagnostic process performed by medics which can obtain a general clinical profile integrating patient information using his scientific knowledge. Available clinical parameters are stored together with genomic information for each patient to create (as much as possible) a complete electronic health record.

Several clinical data that are contained in the electronic health records are related to colon cancer (Read & Kodner, 1999): age, diet, obesity, diabetes, physical inactivity, smoking, heavy alcohol consumption, previous colon cancer or other cancers, adenomatous polyps which are the small growths on the inner wall of the colon and rectum; in most cases, the colon polyp is benign (harmless). Also, other diseases or syndromes such as inflammatory bowel disease, Zollinger-Ellison syndrome, and Gardner's syndrome are related to colon cancer.

In the context of genomic data related with colon cancer, malignant changes of the large bowel epithelium are caused by mutations of specific genes among which we can differentiate (Houlston & Tomlinson, 1997):

- **Protooncogenes.** The most popular mutated protooncogenes in colon cancer are: K-RAS, HER-2, EGFR and c-MYC.
- **Suppressor genes-anticogenes.** In colorectal cancer the most important are DCC, TP53 and APC.
- **Mutator genes.** So far, six repair genes of incorrectly paired up bases were cloned from humans. Four are related to Hereditary Nonpolyposis Colon Cancer (HNPCC). These are: the hMSH2- homolog of yeast gene MutS, the hMLH1 - homolog of bacterial MutL, the hPMS1 and hPMS2 - from yeast equivalent.

An efficient way to process the above gene sequences is to detect Single Nucleotide Polymorphisms (SNPs) (Sielinski, 2005). SNPs data are qualitative data providing information about the genomic at a specific locus of a gene. An SNP is a point mutation present in at least 1% of a population. A point mutation is a substitution of one base pair or a deletion, which means the respective base pair is missing or an addition of one base pair. Though several different sequence variants may occur at each considered locus, usually one specific variant of the most common sequence is found, an exchange from adenine (A) to guanine (G), for instance. Thus, information is basically given in the form of categories denoting the combinations of base pairs for the two chromosomes, for example, A/A, A/G, G/G, if the most frequent variant is adenine and the single nucleotide polymorphism is an exchange from adenine to guanine.

According to previous medical knowledge, there are several SNPs with known relation to colon cancer. Some indicative SNPs already related to colon cancer, according to several sources in the literature, identified in TP53 gene, are presented in Table 1. The expert knowledge contains information about the position of the SNPs in the gene sequence (i.e., exon, codon position and amino acid position), the transition of the nucleotides and the translation of the mRNA to protein. Based on the list of known SNPs related to colon cancer, appropriate genomic information is derived, revealing the existence, or not, of these SNPs in the patient's genes.

Some of the described genes are acquired from the subjects and based on the SNP information regarding every acquired gene, such as SNPs in Table 1 for TP53 gene, new features are derived. These new features contain information regarding the existence or not of these SNPs in the patient's gene sequence. The derived features along with the aforementioned clinical data that are related to colon cancer are the input to the methodology and, after following the above described inference engine creation methodology, clinicogenomic profiles are generated. These profiles are able to provide advanced cancer decision support to new patients.

FUTURE TRENDS

There should be no doubt that several challenges remain regarding clinical and genomic data integration

Table 1. Indicative SNPs transitions and positions in the TP53 gene, related with colon cancer

Region	mRNA pos.	Codon pos.	Amino acid pos.	Function	Transition	Protein residue transition
exon_10	1347	1	366	nonsynonymous	G/T	Ala [A]/Ser [S]
exon_10	1266	1	339	nonsynonymous	A/G	Lys [K]/Glu[E]
exon_9	1242	3	331	synonymous	A/G	Gln [Q]/Gln[Q]
exon_8	1095	1	282	nonsynonymous	T/C	Trp [W]/Arg[R]
exon_8	1083	1	278	nonsynonymous	G/C	Ala [A]/Pro[P]
exon_8	1069	2	273	nonsynonymous	A/G	His [H]/Arg[R]
exon_7	1021	2	257	nonsynonymous	A/T	Gln [Q]/Leu[L]
exon_7	998	3	249	nonsynonymous	T/G	Ser [S]/Arg[R]
exon_7	994	2	248	nonsynonymous	A/G	Gln [Q]/Arg[R]
exon_7	984	1	245	nonsynonymous	A/G	Ser [S]/Gly[G]
exon_7	982	2	244	nonsynonymous	A/G	Asp [D]/Gly[G]
exon_7	973	2	241	nonsynonymous	T/C	Phe [F]/Gly[G]
exon_5	775	2	175	nonsynonymous	A/G	His [H]/Arg[R]
exon_5	702	1	151	nonsynonymous	A/T/C	Thr [T]/Ser[S]/Pro [P]
exon_5	663	1	138	nonsynonymous	C/G	Pro [P]/Ala [A]
exon_5	649	2	133	nonsynonymous	C/T	Thr [T]/Met [M]
exon_4	580	2	110	nonsynonymous	T/G	Leu [L]/Arg [R]
exon_4	466	2	72	nonsynonymous	G/C	Arg [R]/Pro [P]
exon_4	390	1	47	nonsynonymous	T/C	Ser [S]/Pro [P]
exon_4	359	3	36	synonymous	A/G	Pro [P]/Pro [P]
exon_4	353	3	34	synonymous	A/C	Pro [P]/Pro [P]
exon_2	314	3	21	synonymous	T/C	Asp [D]/Asp [D]

to facilitate clinical decision support. The opportunities of combining these two types of data are obvious, as they allow obtaining new insights concerning diagnosis, prognosis, and treatment. According to this, medical informatics are combined with bioinformatics towards biomedical informatics. Biomedical Informatics is the emerging discipline that aims to put these two worlds together so that the discovery and creation of novel diagnostic and therapeutic methods is fostered. A limitation of this combination is that although data exist, usually their enormous volume and their heterogeneity constitute their analysis and association a very difficult task. Another challenge is the lack of terminological and ontological compatibility, which could be solved by means of a uniformed representation. Besides new data models, ontologies are/have to be developed in order to link genomic and clinical data. Furthermore, standards are required to ensure interoperability between disparate data sources.

CONCLUSION

Advances in genome technology are playing a growing role in medicine and health care. With the development of new technologies and opportunities for large-scale analysis of the genome, genomic data have a clear impact on medicine. Cancer prognostics and therapeutics are among the first major test cases for genomic medicine, given that all types of cancer are related with genomic instability. The integration of clinical and genomic data makes the prospect for developing personalized health care ever more realistic.

REFERENCES

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., et al. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Cell Biology, 96*, 6745-6750.

Bezdek, J.C. (1981). Pattern recognition with fuzzy objective function algorithms. Plenum Press, USA.

Boon, M.E., & Kok, L.P. (2001). *Using artificial neural networks to screen cervical smears: How new technology enhances health care*. Clinical applications of artificial neural networks. Cambridge: Cambridge University Press (pp. 81-89).

Chiang, I.J., Shieh, M.J., Hsu, J.Y., & Wong, J.M. (2005). Building a medical decision support system for colon polyp screening by using fuzzy classification trees. *Applied Intelligence, 22*(1), 61-75.

Cook, D.J., Lawrence, B.H., Su, S., Maglothin, R., & Jonyer, I. (2001). Structural mining of molecular biology data: A graph-based tool for discovering and analyzing biological patterns in structural databases. *IEEE Engineering in Medicine and Biology, 4*, 67-74.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*(1), 1-38.

Doornewaard, H., van der Schouw, Y.T., van der Graaf, Y., Bos, A.B., Habbema, J.D., & van den Tweel, J.G. (1999). The diagnostic value of computerassisted primary cervical smear screening: A longitudinal cohort study. *Modern Pathology, 12*(11), 995-1000.

Exarchos, T.P., Papaloukas, C., Lampros, C., & Fotiadis, D.I. (2006). Protein classification using sequential pattern mining. In *Proceedings of the 28th Annual International Conference of the Engineering in Medicine and Biology Society,* USA (pp. 5814-5817).

Fotiadis, D.I., Goletsis, Y., Likas, A., & Papadopoulos, A. (2006). *Clinical decision support systems*. Encyclopedia of Biomedical Engineering, Wiley.

Houlston, R.S., & Tomlinson, I.P.M. (1997). Genetic prognostic markers in colorectal cancer. *Journal Clinical Pathology: Molecular Pathology, 50*, 281-288.

Jiang, X.R., & Gruenwald, L. (2005). Microarray gene expression data association rules mining based on BSC-tree and FIS-tree. *Data & Knowledge Engineering, 53*(1), 3-29.

Louie, B., Mork, P., Martin-Sanchez, F., Halevy, A., & Tarczy-Hornoch, P. (2007). Data integration and genomic medicine. *Journal of Biomedical Informatics, 40*, 5-16.

MacQueen, J.B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281-297).

Mitchell, T. (2006). *Machine learning*. Springer, McGraw-Hill Education (ISE Editions).

Nieminen, P., Hakama, M., Viikki, M., Tarkkanen, J., & Anttila, A. (2003). Prospective and randomised public-health trial on neural network-assisted screening for cervical cancer in Finland: Results of the first year. *International Journal of Cancer, 103*(3), 422-426.

Parekattil, S.J., Fisher, H.A., & Kogan, B.A. (2003). Neural network using combined urine nuclear matrix protein-22, monocyte chemoattractant protein-1 and urinary intercellular adhesion molecule-1 to detect bladder cancer. *The Journal of Urology, 169*(3), 917-920.

Read, T.E., & Kodner, I.J. (1999). Colorectal cancer: risk factors and recommendations for early detection. *American Family Physician, 59*(11), 3083-3092.

Remzi, M., Anagnostou, T., Ravery, V., Zlotta, A., Stephan, C., Marberger, M., et al. (2003). An artificial neural network to predict the outcome of repeat prostate biopsies. *Urology, 62*(3), 456-460.

Shah, S., & Kusiak, A. (2007). Cancer gene search with data-mining and genetic algorithms. *Computers in Biology and Medicine, 37*(2), 251-261.

Sielinski, S. (2005). *Similarity measures for clustering SNP and epidemiological data.* Technical report, University of Dortmund.

Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., et al. (2002). (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell, 1*(2), 203-209.

Skevofilakas, M.T., Nikita, K.S., Templakesis, P.H., Birbas., K.N., Kaklamanos, I.G., & Bonatsos, G.N. (2005). A decision support system for breast cancer treatment based on data mining technologies and clinical practice guidelines. In *Proceedings of the 27th Annual Conference of IEEE Engineering in Medicine and Biology* (pp. 2429-2432), China.

Slaymaker, M., Brady, M., Reddington, F., Simpson, A., Gavaghan, D., & Quirke, P. (2006). A prototype infrastructure for the secure aggregation of imaging and pathology data for colorectal cancer care. In *Proceeding of IEEE Computer Based Medical Systems* (pp. 63-68).

Tan, P.N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Addison Wesley. USA.

Van'T Veer, L.J., Dai, H., Van De Vijner, M.J., He, Y.D., Hart, A.A.M., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature, 415*, 530-536.

Walker, P.R., Smith, B, Liu, Q.Y., Famili, A.F., Valdes, J.J., Liu, Z., et al. (2004). Data mining of gene expression changes in Alzheimer brain. *Artificial Intelligence in Medicine*, *31*(2), 137-154.

Zaki, M.J. (2000). Sequence mining in categorical domains: Incorporating constraints. In *Proceedings of the 9th international conference on information and knowledge management,* USA (pp. 422-429).

KEY TERMS

Cancer Staging: Knowledge of the extent of the cancer at the time of diagnosis. It is based on three components: the size or depth of penetration of the tumour (T), the involvement of lymph nodes (N), and the presence or absence of metastases (M).

Clinical Decision Support Systems (CDSS): Entities that intend to support clinical personnel in medical decision-making tasks. In more technical terms, CDSSs are active knowledge systems that use two or more items of patient data to generate case-specific advice.

Cluster analysis: The task of decomposing or partitioning a dataset into groups so that the points in one group are similar to each other and are as different as possible from the points in other groups.

Data Integration: The problem of combining data residing at different sources and providing the user with a unified view of these data. This important problem emerges in several scientific domains, for example, combining results from different bioinformatics repositories.

Data Mining: The analysis of observational datasets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.

Mutation: A change in the genetic material (usually DNA or RNA) of a living being. Mutations can happen for a lot of different reasons. They can happen because of errors during cell division, because of radiation, chemicals, and so forth.

Single Nucleotide Polymorphism (SNP): A DNA sequence variation occurring when a single nucleotide—A, T, C, or G—in the genome differs between members of a species or between paired chromosomes in an individual.

Tumour Suppressor Gene: A gene that reduces the probability that a cell in a multicellular organism will turn into a tumor cell. A mutation or deletion of such a gene will increase the probability of the formation of a tumor.

An Intelligent Wearable Platform for Real Time Pilot's Health Telemonitoring

Christos Papadelis
Greek Aerospace Medical Association and Space Research, Thessaloniki, Greece

Chrysoula Kourtidou-Papadeli
Greek Aerospace Medical Association and Space Research, Thessaloniki, Greece

Fotini Lazaridou
Greek Aerospace Medical Association and Space Research, Thessaloniki, Greece

Eleni Perantoni
Greek Aerospace Medical Association and Space Research, Thessaloniki, Greece

INTRODUCTION

Aviators engage in a variety of outdoor activities where their health status, the environment, and the degree of workload and fatigue affect their performance. An innovative tool has been developed, which supports the real-time health monitoring of pilots using new algorithms based on intelligent clustering techniques for the recognition of possible health problems in flight. The Smart Profiler and the Intelligent Advisor modules of this system exploit the use of knowledge based expert systems and intelligent classification techniques. Coupled with the Portal, which also exploits the use of intelligent clustering techniques, it estimates the pilot's performance in unknown environments. The new system targets recognizing possible problems at the time of flying, but it can also be used for the monitoring of the pilot performance and progress throughout a period of time, as it stores information from different flying sessions. The system was applied in 20 private pilots during the flight of a Cessna 152 aerobatic. The device was reliable and user-friendly, enabling us to monitor real-time health status of aviators in order to detect possible problems caused by the actual environmental conditions to which individuals are exposed, thus contributing to their health and safety in their working environments.

Despite the automation and increasing technological complexity of modern aircrafts, the human operator still plays an important role in controlling those demanding systems. Piloting an aircraft is a highly complex task that requires the pilot to be proficient in numerous skills (Wilson & Eggemeier, 1991) in a hostile environment of cabin pressure changes and circadian rhythm disturbances particularly in long duration flights. The resulting overload of the pilots mandates the need for real time health telemonitoring (Charles, Winget, Charles, DeRoshia, Markley, & Holley, 1984; Denison, Ledwith, & Poulton, 1966; U.S. National Research Council, 2002; Ustinavičienė, Obelenis, & Ereminas, 2004). Real time health telemonitoring would be crucial to early detect and prevent conditions affecting aviator's vital signs and cognitive performance.

BACKGROUND

A series of technologically advanced devices has been gradually proposed in order to detect pilot performance decrement, but most of the methods referred to pre or post flight data analysis and basically in simulation flights. Most of the commercially available wearable or portable devices measuring health parameters were mainly developed in sports activities for athletes. Initially those products did not support wireless transmission of data for long distances. Long-distance measurements were made possible due to newer products such as Cardiosport (6), FitSense (7), Polar S-Series Monitors (8), and FS-1 speedometer; which share the drawbacks of limited storing capacity; only one parameter monitoring and no wireless data transmission support (or when provided, only for very short distances). Newer products as MySportTraining (9), WebCoach (10), CrossTrak 2.0 and CoachConnect (11) and the Virtual Coach (12) offered the possibility of calculating personalized indicators, but none of these products, however, record or use information related to the environmental conditions of the training place. Recently a product in the form of life shirt was presented by VivoMetrics™-LifeShirt (13) embedded with tiny wires and electrodes which continuously monitor 40 physiological signs of sickness and health. The problem with this device included high cost and the insufficient adaptability to individual's body features. Additionally, it did not incorporate capabilities of Computer Aided Diagnosis so the continuous presence of a doctor was mandatory. In all those products, environmental parameters were not included as it is recommended in the case of pilot's activities.

AN INTELLIGENT WEARABLE PLATFORM

The goal of the present study was to assist in the implementation and the application of an innovative device for real time health telemonitoring to early recognise potential hazards and alert the pilot or the supervising personnel, using a neural network technique, which enables the system to incorporate adaptations of the real environment, and estimate possible future events that may occur during an intended flight in similar environment.

The proposed system consists of two architectures designed, one for monitoring a group of pilots and the other for a pilot alone (standalone). The difference between those architectures is that for the standalone version, the wearable unit is embedded inside the Monitoring Station while for a group of pilots it is separate and on the ground.

The system for a group of pilots is more complicated and will be analysed here. This system is composed of a set of sensors distributed on the body via a wearable comfortable vest, underneath the pilots clothes, connected to a portable wearing device where the communication module is incorporated receiving information from three sources: the environment and the medical and training profile of the pilots. Through a SIM card which contains a microprocessor chip which stores unique information about an account and identifies that to the network, all the data are transmitted online to the monitoring station where the processing and analysis is accomplished until the final recognition of the problem when the appropriate alert will appear on the screen of the monitoring personnel. Independently of this system, another external monitoring system, the portal, is used to keep a record of the pilot performance in each flight and through the intelligent clustering mechanism. This is compared to similar situations. This whole process is succeeded by seven subsystems: more analytically

Figure 1. The system overview

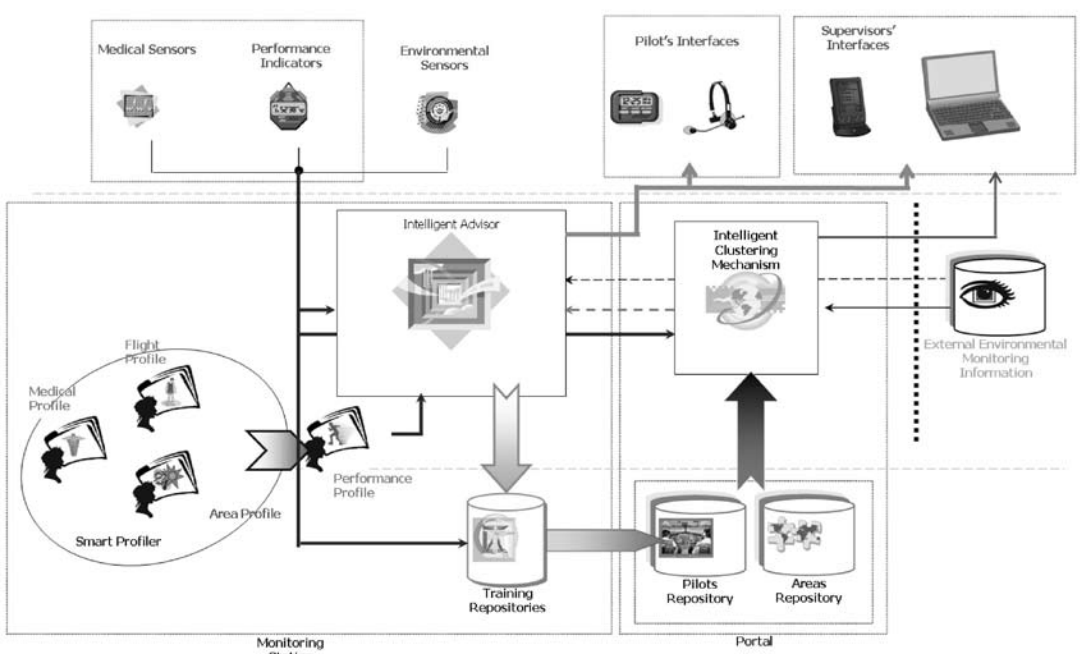
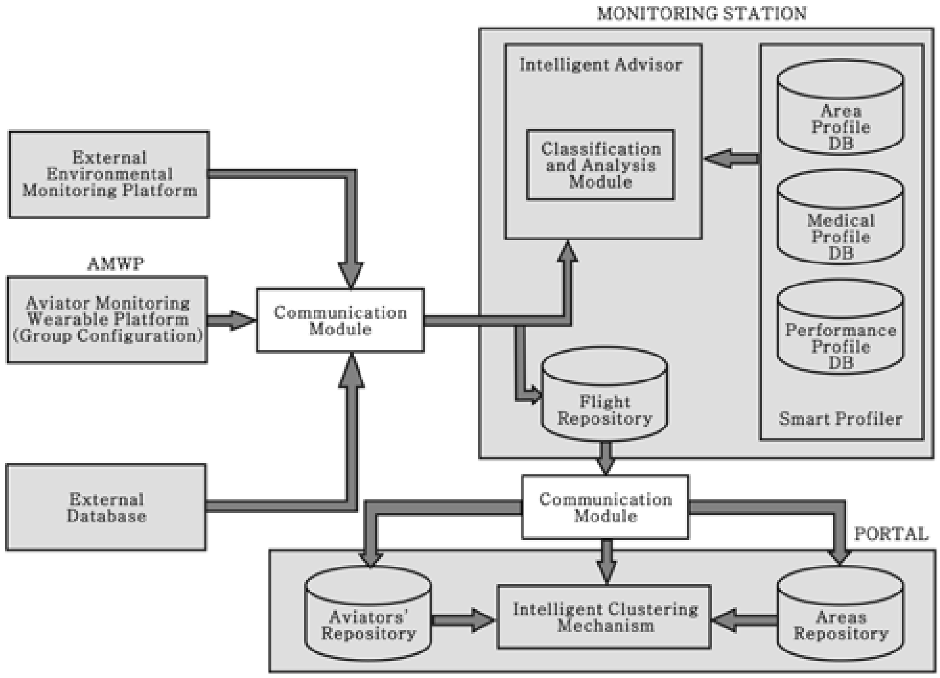


Figure 2. The architecture of the full functionality Configuration PREVENTIVE system



the Aviator Monitoring Wearable Platform (AMWP), the Monitoring Station, the Portal, the Interface to external Environmental Monitoring System, the Interface to external Training Database, the Communication Module, and the Security Module.

More specifically, the AMWP, which is physically attached on the aviator's body with a set of sensors, is composed of the following modules: medical sensing system, which acquires and digitizes medical measurements, performance sensing system, which acquires and digitizes performance measurements, and integrator, which synchronizes measurements from both the medical and performance sensing systems. The hardware part of the WAD unit consists of all the internal circuitry and the communication interfaces. Figure 3 depicts the block diagram of the WAD.

For the sensing systems, the sensors are distributed as follows: two for respiration, one for temperature, one for pulse oximeter (NONIN 8000J, Nonin Medical Inc.), one for wrist blood pressure meter (OMRON 637IT, Omron Healthcare Europe), one for 5-lead ECG, one for accelerometer, and all were embedded in the WAD unit. The signals were acquired using analogue interface for ECG, respiratory, temperature, and acceleration sensors and digital for the pulse oximeter (P-OX) and blood pressure meter (BPM). Each portable unit was supplied with a preconfigured GPRS-enabled wearable device and any data collected were transmitted via a GPRS network to a Gateway server with an IP address publicly accessible. In turn, the Gateway server stored the data to a database server where all data reside and a visualization subsystem which typically includes a PDA for viewing the results of the performed tests through wireless connection to the Web server, enabling specialized personnel to real time monitor the aviator through the Monitoring Station.

The Monitoring Station is a portable computer consisting of three software modules to process all incoming data: the Smart Profiler to generate the performance profile, using information from the medical profile, area, and flight profile which are processed, analyzed, and calcified by the intelligent advisor and stored in the Training Repository. The Smart Profiler works by using personalized information to specify each pilot's performance profile, which expresses the pilot's capacity for the specific environmental conditions, medical history, current health status, and flight requirements. It is formulated from the intelligent incorporation of the pilot's Medical Profile, the Flight Profile, and the Area Profile. The three profiles can either be created automatically, or manually. The medical profile changes as pilots are submitted to frequent medical examinations,

Figure 3. The Smart Profiler Expert System

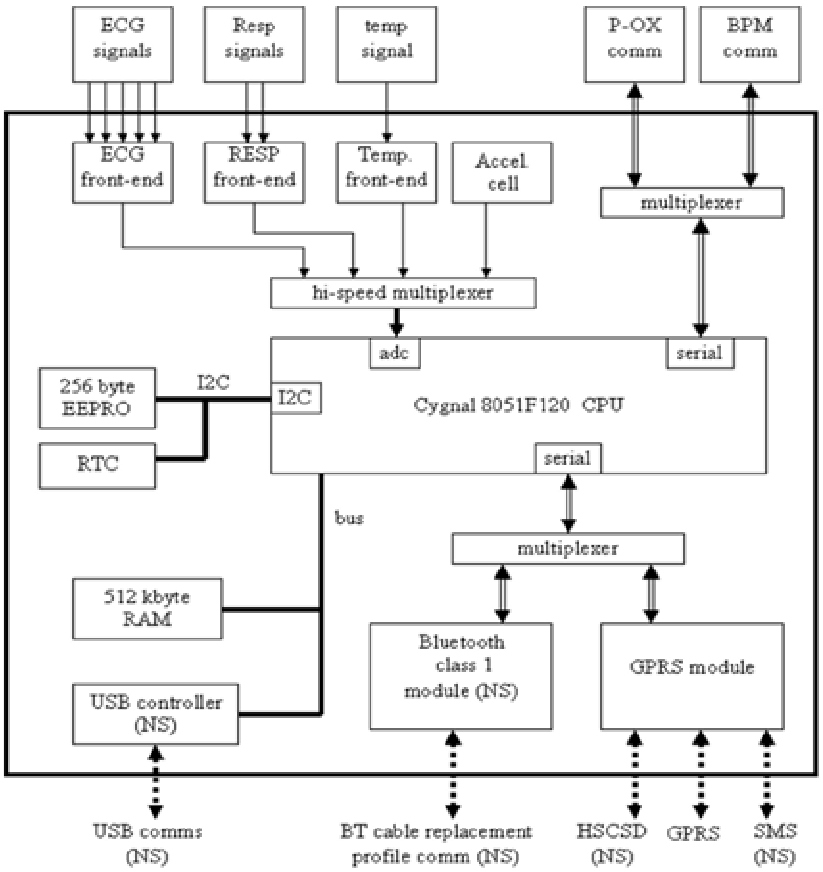
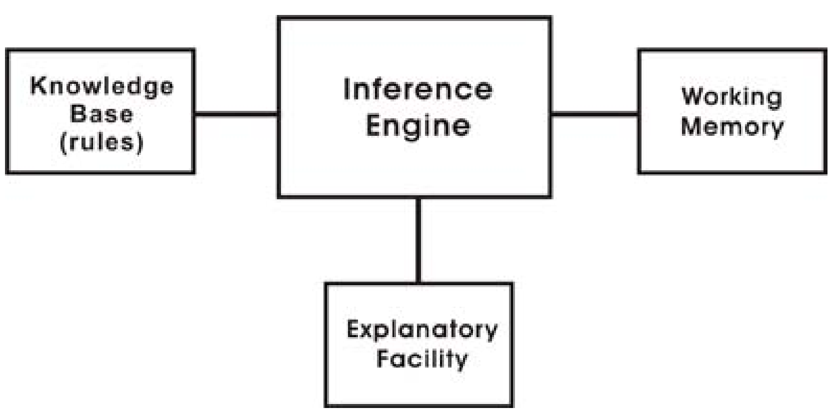


Figure 4. Block diagram of the WAD unit



I

which provide new information on their health status. Flight profiles record the existing knowledge on how the body reacts to certain types of flight. Area profiles exploit the recorded environmental monitoring parameters for a region. They are strongly time-dependant and change dynamically as more monitoring measurements and data become available. The Smart Profiler, which is actually a knowledge-based expert system through one of its components, the working memory, receives the three basic profiles of a pilot who is about to fly in a specific location, and uses the already recorded knowledge from previous similar training sessions to provide an estimation of the pilot's performance and his vital response during this specific flying session. In order to achieve this, a knowledge base is used which is actually a set of rules which apply to the facts of the expert system and can be seen as the codification of the experts' knowledge from the medical and flight domain. Then a managing mechanism is used, called inference engine, which is capable of deciding which rules are satisfied by facts, prioritizing the satisfied rules, and executing the rule with the highest priority. Finally, the explanatory facility explains the reasoning of the final result (the performance profile) to the end users.

Besides the smart profiler, the second component of the Monitoring Station is the Intelligent Advisor, which processes real-time monitoring data from the medical, environmental, and performance sensors, combining them with the smart profiles in order to classify the aviator's status according to the significance of a possible abnormality. An innovative technique that the system is using is the clustering algorithm in order to process data by classifying flight areas in terms of both performance capabilities and associated health risks. Those areas are represented in high-dimensional spaces and hence the inherent similarities are hard to recognize and to illustrate. In order to facilitate this process, a neural network model that is capable of visualizing these similarities was built, based on the Self Organising Map (SOM) algorithm. The Intelligent Clustering Mechanism, by using the SOM algorithm, creates clusters of areas according to the similarities that the areas present and for each area two different estimations are provided, one for the performance and one for the aviator's health response. The SOM algorithm, in order to create the clusters of areas, needs as input the estimated importance factor of each of the above parameters, which is subjected to changes according to the type of flight. The estimation for the possible health problems that may appear in the aviator's health is also provided for every flight area that belongs to a specific cluster.

The third component of the Monitoring Station is the Flight Training Repository, which stores information about a flight session for each aviator (environmental conditions of the flight area, as well as medical and performance parameters); and provides diagnosis and advices generated from the Intelligent Advisor. Information from the Flight Repository is send through the communication module to another external monitoring system, the Portal.

The Portal is an independent part of the system, the purpose of which is to provide the long-term service of gathering results from a large number of training sessions throughout different locations and time periods, and extract conclusions and predictions on the future course of a pilot. Consequently, it is a self-learning and maintenance mechanism that gets improved as time goes by: the more the training sessions in a specific area are, the more accurate the produced area similarity maps become. Therefore, its results and usefulness can be evaluated only after a long period of time and after the insertion of a great number of a data. Another advantage that the portal adds to the platform is centralization. Having a head repository that keeps all training area data to the portal, we offer/ we present a unified system

Table 1. The importance factor of each environmental parameter according to the flight requirements

Areas clustering with SOM algorithm							
Altitude	Importance factors of each parameter						
	WBGT	O_3	CO	SO_2	NO_2	Pm	Altitude
Height<= 500 m	80		5	5	5	5	
500 m < Height <= 1000 m	15	5					80
Height >1000 m	5	5					90

that prevents duplication of records in the monitoring stations. All area attributes are kept inside one repository and training information about an area can be gathered from multiple resources independently, without confusing instances. The Preventive Portal user interfaces are implemented in simple HTML, in conjunction with ASP.NET Server Controls, and VB.NET programming for the site's dynamic parts.

The data from the environmental conditions are entered through an external subsystem, the Interface to an External Environmental Monitoring System which is used to obtain environmental measurements and import them to the platform through the communication module in order to form the profile of the training area over time and to evaluate how this may affect the aviator's performance. The communication model is also receiving training information from another subsystem, the Interface to an External Training Database, which provides the platform with medical and training data for an aviator in order to form his medical and training profile over time.

Overall, the Communication Module is the medium for transferring data between the subsystems and consists of a serial communication between the Configurator application and the WAD unit, communication of the sensors with the WAD unit, and a wireless communication consisting of GPRS connection between the WAD unit and the Gateway application.

In order to configure all the three submodules, the wireless communication with a remote Monitoring Station and the set-up of a training session the software component of the Wearable Platform, the WTA Suite, was used. The applications included in the software were the Training Assistant:, the Configurator, and the Gateway.

The Training Assistant is a .NET application allowing a user to remotely connect to a wearable device, to set all the acquisition session parameters, to start/stop an acquisition session, and to view and record all of the acquired data in realtime through a series of graphs and fields.

The Configurator is a .NET application suitable to configure a wearable device through a serial cable attached to a serial port, through which all communication and security related parameters of the wearable device can be entered. The WAD unit includes a subscriber identity module (SIM) card with activated general packet radio service (GPRS) communication subsystem, which is a connectivity solution based on Internet Protocols, that transmits the acquired measurements to a Web server, and a visualization subsystem which typically includes a personal digital assistant (PDA), which is a handheld device that combines computing, telephone/fax, Internet, and networking features and is used for viewing the results of the performed tests.

Because of the personal information transmitted through the communication module, the existence of a Security Module ensures the overall security within the platform and stores personal data of the user related to demographics, medical condition, and medical history. Security is implemented along the communication paths between servers and other devices. Servers security refers to the protection of devices hosting PREVENTIVE's applications, basic set-up security which refers to basic software configuration that is almost standard when implementing a security policy, and application security which refers to the security mechanisms employed in the application code or which are specific to the PREVENTIVE application apart from those applied to the communication channels.

In case of emergency situations, the advice or alert is communicated either to the aviator through optical or acoustic messages or to the supervising personnel, or both if requested. The form of the messages is simple and easily recognizable, without distracting the pilot from the engaged activity.

The completed system was tested in 20 private pilots of approximately 150 flying hours experience (16 men with a mean body weight of 78.6±9.6 kg, and 4 women with a mean body weight of 72±14.1 kg)

Table 2. The mean values of the parameters compared before and during flight

Measurements	Before Flight	During Flight
Systolic Blood Pressure (mmHg)	122.6 ± 10.6	139.4 ± 12.2
Diastolic Blood Pressure (mmHg)	72.0 ± 6.8	72.7 ± 6.7
Heart rate	75.3 ± 7.5	91.3 ± 10.1
Oxygen Saturation (SaO_2)	97.5 ± 1.0	87.7 ± 0.9

during a baseline session and during flight manoeuvres in an altitude of 0 – 10000 feet. Pilots mean age was 34.8±14.2 years for men and 37.5±11.7 years for women. They all signed an informed consent form. The airplane used was a Cessna 152 aerobatic. The flight profile was start up engine, taxiing, take off, climb, level flight, stall manoeuvre at 5.000 feet, an ascent to 10.000 feet with a negative g maneuver , and then a positive g maneuver to descend in forced landing with no touch procedure but go around for a final landing to the airport. The total session lasted 1 hour and 30 minutes. It was repeated once a week for 4 consecutive weeks.

The data flow was continuous with a delay of a few seconds. After each testing session, a measurements datasheet was filled by the medical doctor with the pilot's data. Questionnaires and interviews were also used for the system ergonomics, usability, friendliness, and so forth. Final comments and useful feedback was gathered from doctors, medical professionals, and pilots.

The first measurements were taken right after the vest was applied and connected to the server. Those measurements served as rest data. After the pilot was settled in the airplane the second measurement was taken as baseline data. Continuous measurements were taken during taxiing, take off, actual flight, and landing. From the actual measurements, generally we revealed differences in heart rate, oxygen saturation, and systolic blood pressure during hypoxia and stress in flight.

There was a considerable inter and intra subject variability in response to different parts of flight which imposes the need for a larger number of subjects. The interest at this stage was given from the scientists to the functionality of the system and the successful recording of the vital signs and performance of the pilots in various flying conditions, and their correlations with advanced algorithms in order to be clustered.

The advantages of the PREVENTIVE device are the monitoring of a wide range of physiological signs, coupled with performance indicators and on-site, environmental parameters, providing real-time, continuous monitoring and online feedback, supporting both individual pilots and groups of pilots, capable of being "packaged" in an unobtrusive, user-friendly wearable device.

The questions asked mainly referred to clinical functionality and more specifically to the operability of the preventive system in a realistic environment, which involves ease of use in a cockpit environment and evaluation of feedback methods. The clinical evaluation was also included in the questionnaire, referring to the evaluation and assessment of the diagnosis module, whether the system was functional, and finally the users were asked to provide comments and suggestions on issues related to usability as comfort, feedback, and incentive for use, overall functionality on best and worst things about the system. Doctors and instructors found both software and hardware of the system very useful and easy to operate with some suggestions for improvement. More specifically the general opinion was that the preventive system contributes to increase the benefit of pilots' health and potential risks during flying.

FUTURE TRENDS

Further investigation is necessary for the development of a wearable platform for real time pilots' health telemonitoring. There is a further plan to improve this product in order to resolve small problems that appeared during the study. One of the problems was the oximeter finger sensor which was annoying, especially during the maneuvers. It was placed to the left hand and attached with a tape. It was suggested that the sensor be replaced by an ear lobe sensor. Artifacts in ECG signal due to body movement, hairy chest, and body sweat seen in the majority of pilots during their first flight, were faced with new patches firmly attached to the skin (that was previously thoroughly cleaned and prepared by soft soap and alcoholic solution). There were suggestions from the users to make the wearable device lighter and smaller in size. In a future version of the product, it may be considered the use of other wireless communication means, memorization of larger amount of data on small modules, acquisition extended to other biomedical signals, implementation of a simple alerting system onboard, upgrades of the software applications, upgrades of the user interface on the WAD unit, and upgrades on the biomedical t-shirt, as well as new accelerometer modules capable of acquiring real time three-axial acceleration measurements.

CONCLUSION

Real time health telemonitoring is crucial for the early detection and prevention of conditions affecting aviators' vital signs and cognitive performance. PREVENTIVE system allows the on-board real time pilot's health telemonitoring, though further improvements are necessary.

REFERENCES

Charles, M., Winget, B.A., Charles, W., DeRoshia, M.A., Markley, C.L., & Holley, D.C. (1984, December). A review of human physiological and performance changes associated with desynchronises of biological rhythms. Science and Technology Committee: Selected Review Paper (Vol. 55, No. 12). Aviation, Space and Environmental Medicine.

Denison, D.M., Ledwith F., & Poulton E.C. (1966). Complex reaction times at simulated cabin altitudes of 5,000 feet and 8,000 feet. *Aerosp Med., 37*(10), 1010-1013.

U.S. National Research Council. (2002). *The airliner cabin environment and the health of passengers and crew.* New York: National Academic Press.

Ustinavičienė, R., Obelenis, V., & Ereminas, D. (2004). Occupational health problems in modern work environment. *Medicina (Kaunas), 40*(9).

Wilson, G.F., & Eggemeier F.T. (1991). Physiological measures of workload in multi-task environments. In D. Damos (Ed.), *Multiple-task performance* (pp. 329-360). London: Taylor and Francis.

Retrieved January 30, 2008, from http://www.cardio-sport.com

Retrieved January 30, 2008, from http://www.fitsense.com

Retrieved January 30, 2008, from http://www.polar.fi

Retrieved January 30, 2008, from www.musporttraining.com

Retrieved January 30, 2008, from www.sailcoach.com

Retrieved January 30, 2008, from www.crosstrak.com

Retrieved January 30, 2008, from www.trainingbible.com

Retrieved January 30, 2008, from www.vivometrics.com

KEY TERMS

Health Telemonitoring System: A system that gives the ability to monitor healthy individuals or chronically ill patients by distance.

Hypoxia: A pathological condition in which the body as a whole or region of the body is deprived of adequate oxygen supply.

Neural Network: A computing solution that is loosely modeled after cortical structures of the brain. It consists of interconnected processing elements called nodes or neurons that work together to produce an output function.

Personal Digital Assistant (PDA): Handheld computers that were originally designed as personal organizers, but which have become much more versatile over the years.

Portal: A site on the World Wide Web that typically provides personalized capabilities to its visitors, providing a pathway to other content.

Subscriber Identity Module (SIM) Card: A removable digital smart card that can store securely the key identifying a mobile phone service subscriber.

Workload: The relationship between a group or individual human operator and task demands.

Interactive Video Game-Based Tool for Dynamic Rehabilitation Movements

Aimee L. Betker
University of Manitoba, Canada

Tony Szturm
University of Manitoba, Canada

Zahra Moussavi
University of Manitoba, Canada

INTRODUCTION

In being seated, standing, and walking, many uncontrollable factors contribute to the degradation of our balance system. The maintenance of balance involves many essential sensory (visual, vestibular, and somatosensory) and motor processes. Each sensory input provides unique internal and external reference frame information to the central nervous system (CNS). The CNS interprets the sensory information, from which preplanned and/or preventative (feedforward controls) and corrective (feedback controls) actions can be taken and conflicting sensory information can be mediated (Peterka, 2002). In the absence of a sensory input, balance can still be maintained; however, the compensatory actions become larger and different balance strategies may be employed.

Serious problems facing older adults and many people with neurological disorders (e.g., stroke, traumatic head injuries, incomplete spinal cord injuries, Parkinson's, multiple sclerosis, diabetic peripheral neuropathy, and osteoarthritis) are balance impairment, mobility restriction, and falling (Gill et al., 2001; Harris, Eng, Marigold, Tokuno, & Louis, 2005). In these cases, even small disturbances may result in a fall and injuries are very likely to occur. This increased risk of falling combined with mobility limitations precipitates patient dependency in instrumental and basic activities of daily living; in turn, this results in reduced levels of physical activity.

Due to the many problems associated with reduced balance and mobility, providing an effective rehabilitation regime is essential to progress recovery from impairments and to help prevent further degradation of motor skills. Thus, we developed a tool which uses a computer-based gaming system to facilitate the rehabilitation process for restoration of weight bearing and balance control, for persons with neurological or musculo-skeletal disorders.

BACKGROUND

There are many aspects to consider when developing therapy programs for balance impairments and mobility limitations. Rehabilitation regimes often include cycling, standing activities, steppers, and over-ground and treadmill walking. Dependent on the type of disorder, these exercises will be performed at a given difficulty level (e.g., rate, tension, or incline settings) for a specified duration (Carr & Shepherd, 1998). The effectiveness of these regimes is proportional to the: (1) instensity and volume of training; and (2) task specificity of the exercises (Carr & Shepherd, 1998; Kwakkel, 2006). Constraint-induced movement therapy (CIMT) is a program that directly addresses these two issues (Marklund & Klassbo, 2006). For example, Marklund and Klassbo (2006) applied CIMT to chronic stroke patients, targeting the lower limb. The knee of the less-affected limb was restricted, requiring the tasks to be performed with the affected limb. Results demonstrated that the subjects showed improved dynamic balance and motor ability. While results of these studies are encouraging, repetitive, time-intensive therapies can become tedious, causing a lack of motivation; in turn, the patient may not complete the rehabilitation program. Thus, an important factor of a successful exercise program is maintaining the patient's interest and motivation (Betker, Szturm, Moussavi, & Nett, 2006; Cogan, Madey, Kaufman, Holmlund, & Bach-y-Rita, 1977).

One method to add motivation is through the selection of exercise tasks that are in themselves fun. For example, Tsang and Hui-Chan (2004) found that learning both Tai Chi and golf improved dynamic standing balance control. However, as these are complex tasks, not all patient populations will be capable of participating; thus, methods must be found which will allow even severely disabled individuals to play and be competitive. A promising example is the incorporation of biofeedback into virtual environments or video games, which are based on conventional therapies (Back-y-Rita et al., 2002; O'Connor et al., 2000; Sveistrup, 2004). Biofeedback augmented training presents a functional, task-oriented signal to the subject in a simplified format in order to enhance movement, weight bearing status, or balance awareness (De Weerdt & Harrison, 1985; Glanz, Klawansky, & Chalmers, 1997; Huang, Wolf, & He, 2006). Both virtual environments and video games cognitively engage the patient when they are training, in fun and challenging tasks; in turn, this can increase practice time, volume, and recovery (Bach-y-Rita et al., 2002; Cunningham & Krishack, 1999). In general, video games have been employed less than virtual environments, as the input systems are less developed (e.g., converting center of pressure signal into a mouse signal).

Another consideration is the availability of the system. In order for the tools to be available to a wide range of clinics (for example low and moderately funded clinics) and for at-home use, the equipment should be easy to use, should not incur a large cost, and should be portable. Another aspect, essential for at-home use, is a method to embed the assessment into the treatment regime. Embedded assessment will also remove performance pressure from the patient and automatically log their practice time and performance. Our interactive video game based tool, which adopts the aforementioned ideas, is described in the following section.

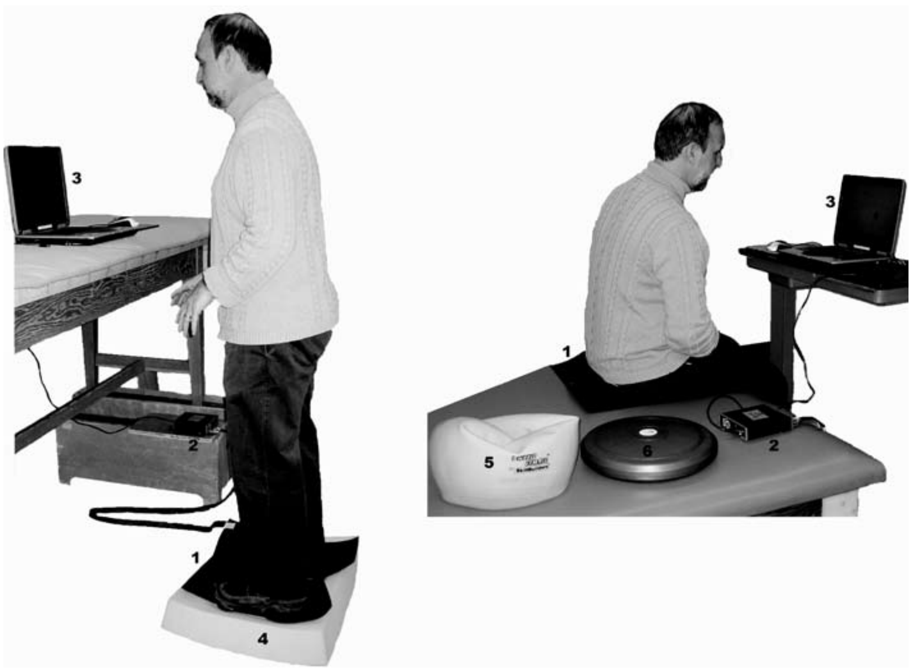
VIDEO GAME-BASED EXERCISE TOOL

Biofeedback Signal Selection and Presentation

The center of pressure (COP) was selected as the biofeedback signal and outcome measure, as it reflects global instabilities of both balance disturbances and reactions (Cheng, Wu, Liaw, Wong, & Tang, 2001; Lee, Wong, & Tang, 1996). For example, in Zambarbieri, Schmid, Magnaghi, Verni, Macellari, and Fadda (1998), insoles were used to determine the COP trajectory; a tactile stimulation device was then used to inform the subject when their COP fell outside an accepted normal range. The amount of load for each leg was displayed on an LED (Cheng et al., 2001; Lee et al., 1996), with a line indicating whether the weight was equally balanced or towards which leg the balance was skewed. In Femery, Moretto, Hespel, and Lensel (2001), a visual display of the footprint sensors was given, with a sound being played if the amount of pressure exceeded a predefined threshold. Similarly, in Urhan and Dincer (2001), the force at the heels was displayed on an LCD, with an audio tone indicating if the correct force is being applied. In Rougier (2004), anterior-posterior (AP) and medial-lateral (ML) displacements of the COP were displayed; different delays were also added to the display to determine their effects on balance. These functional associations between the COP and biofeedback strengthened and created the awareness of the tasks and performance levels, for both the patient and clinician.

As previously mentioned, the COP biofeedback is incorporated into video game-based exercises. A flexible pressure mat (Verg Inc., Winnipeg, MB, Canada) is used to determine the user's COP, which the video game software acquires from the pressure mat's interface box (Figure 1). The COP coordinate acts as the game cursor, emulating either a mouse (analog input) or a joystick (digital input). This is done by mapping the physical COP position signal into a corresponding on-screen pixel coordinate. In order to ensure that the games fully exercise the user's full range of movement, a method is provided to dynamically determine the user's movement range. First, the center point coordinate is determined, that is, the user's overall pressure center point. Next, the peaks of self-induced oscillatory movements are found. These ranges are then displayed on the screen, indicating the portion of the pressure mat which can be activated by the user. As appropriate, the maximal range values can be scaled by a percentage value and the center point coordinates may be offset. For example, if a user's stance was asymmetric, the center value could be offset to try and bias the pressure towards the paretic leg. In order to allow individuals with even small movement ranges to play, only the physical COP positions within the determined move-

Figure 1. System setup: (1) subject stands/sits on the pressure mat, which is connected via (2) the interface box to (3) the laptop. The laptop currently displays the game balloon burst. Example surfaces the pressure mat may be placed on top of are (4) a foam pad, (5) a PhysioGymnic ball, or (6) a SwisDisk



ment range are scaled to the on-screen COP values. As the physical positions are scaled, the resulting on-screen movements may appear jittery; thus, the on-screen coordinates can be smoothed over a given number of samples to reduce this effect.

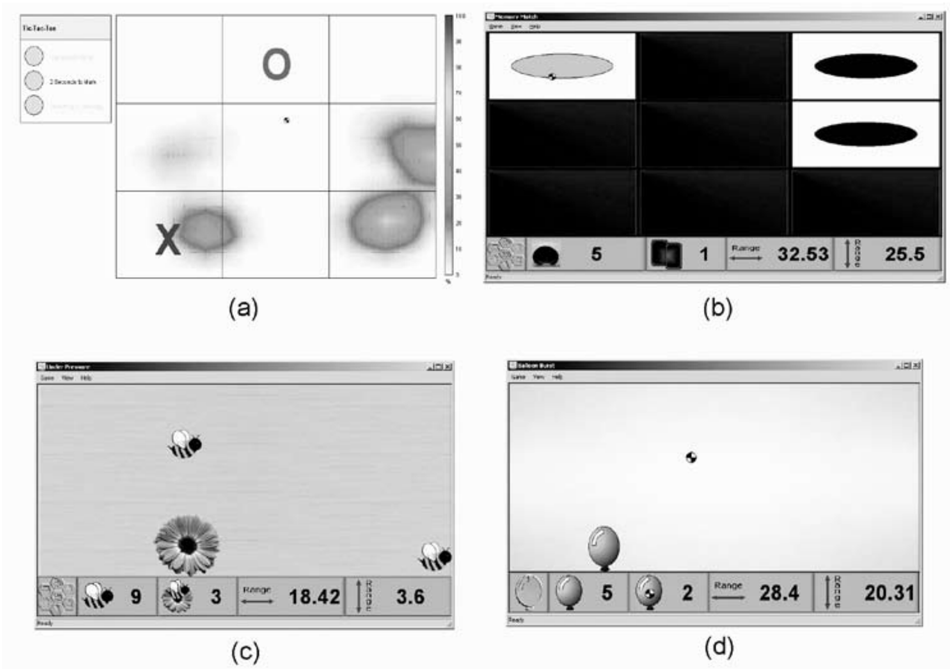
Video Game-Based Tool

The following paragraphs describe game play for each of the four games, Tic-Tac-Toe, Memory Match, Under Pressure, and Balloon Burst (Figure 2).

In Tic-Tac-Toe, the COP coordinate emulates a joystick controller. The screen is divided into nine squares and the user must shift their weight in order to move the on-screen COP cursor to the desired square. The goal is to create a line, with the computer serving as the opponent. The user is instructed what to do via a display panel: (Red) the user may rest, as it is the computer's turn; (Yellow) the user should move the on-screen COP cursor to the desired square they wish to occupy. The user has a limited number of seconds with which to mark the square. This value is configurable, allowing the user/physiotherapist to select the appropriate exercise protocol (e.g., can focus on quick or sustained movements). The remaining number of seconds the user has left to move the cursor is indicated to the user. A sound is played for each elapsed second while the user is selecting a square; and (Green) the square marked by the on-screen COP cursor is now marked. If the user selects a square that has previously been marked, the game returns to the Yellow state.

In Memory Match, the COP coordinate emulates a joystick controller. The screen is divided into squares and the user must shift their weight in order to move the on-screen COP cursor to the desired square. The goal is to select two matching cards from a 3-by-3 or 4-by-4 array of squares. Once the COP is held still in a square for a user selected duration, the card is revealed. The second card is then selected in a similar manner; if the cards match, they remain face up. This process is repeated until all of the card pairs are selected. The Memory Match display is similar to that for Tic-Tac-Toe, showing the number of pairs found and the indicators for what stage the game is at: (Red) the two selected cards are checked for a match. When a pair is found, a sound plays and the score is updated; (Yellow) the user should move the COP cursor to the card they would like to turn over. The number of seconds they have left to move the cursor is also indicated. A sound is played for each elapsed second while the user

Figure 2. Screenshots: (a) Tic-Tac-Toe: the user has 2 more seconds to select their square, as indicated by the yellow light; (b) Memory Match: the number of pairs found and the ranges are displayed; (c)Under Pressure: the total number of bees, number of bees caught, and ranges are displayed; and (d) Balloon Burst: the total number of balloons, number of balloons popped, and ranges are displayed



is selecting a card; and (Green) the card that the COP cursor is on is turned face up. If the user selects a card that has previously been selected, the game returns to the Yellow state. Difficulty levels are configurable through: (1) the number of seconds the user has to select a card; and (2) the number of cards displayed: 9 or 16. A cognitive difficulty is added when 9 squares is selected, as 1 card will be without a pair. When the number of cards is 16, the area the COP must be in to select the card is smaller and thus the COP movement must be more precise.

In Under Pressure, the COP coordinate emulates a mouse. The user shifts their weight to move a receptacle in order to catch an object, in one of three modes: (1) horizontal: the user must shift weight side to side; (2) vertical: the user must shift weight back and forth; and (3) both: the user is required to shift weight in all directions. Difficulty levels are configurable through: (1) the receptacle size; (2) the object speed; (3) the number of objects; and (4) the option of multiple objects appearing at a specified interval. The number of objects caught and the total number of objects that have appeared are displayed to the user. The game is over when the number of objects reaches a prespecified maximum number.

In Balloon Burst, the COP coordinate emulates a mouse. The user must shift weight in order to move the on-screen COP cursor over the balloon, with the goal being to pop the balloon. The stationary balloons appear at random or predetermined locations. Difficulty levels are configurable through: (1) the size of the balloon; (2) the duration of the balloon appearance; and (3) a rotation of the on-screen COP cursor, which forces the user to adapt a new motor program (Jones, Wessberg, & Vallbo, 2001). In addition, Balloon Burst includes an embedded assessment feature. When the game is played, an assessment file will be automatically generated and named using an ID and the date and time when the game was played. The assessment file contains the positions of the on-screen COP coordinate, the balloon coordinate, and whether the balloon was popped. From these values, the following parameters can be calculated: (1) the COP's trajectory profile; (2) the time taken to pop the balloon; (3) the trajectory of the COP cursor from the time the balloon appears until when the balloon is popped. This will provide an account of spatial accuracy, indicating endpoint undershoot error and its absolute magnitude and endpoint overshoot error and its absolute magnitude.

System Benefits

Our interactive video game-based exercises offer the following benefits: (1) the elicited COP movements are goal directed, to random or predetermined target locations; (2) the ability to map small movement ranges onto on-screen pixel locations, allowing even severely disabled individuals to play and be competitive; (3) user selected game speeds elicit various movement speeds; (4) configurable game difficulty levels; (5) the games require the users to multitask gaze control (head and smooth pursuit), attention to game play strategy (target motion and predicting final location), and balance control; (6) rewarding goal achievement and positive reinforcement; (7) using the flexible pressure mat allows the games to be performed on firm, fixed surfaces or compliant/irregular surfaces. This adds increased balance requirements as appropriate; and (8) embedded assessment with the game Balloon Burst. In addition, results from case studies with seated and standing subjects indicated that our video-game based exercises were enjoyable and improved participants fall rate performance post-therapy (Betker, Szturm, Moussavi, & Nett, 2006).

A similar approach was taken by NeuroCom NeuroGames, which provides three games that are used to enhance practice and motivation. The center of gravity (COG) is used as the control variable and is measured via a custom strain-gauge force plate. In comparison to NeuroGames, our system offer several advantages. First, any required item selection is done independently of external aids, through a configurable time duration parameter. In NeruoGames Puzzle Master and Solitaire, the user is required to use a mouse must be used to select the desired object. The use of a mouse is undesired; in some cases the user may not be able to operate a mouse; the operation of a mouse provides an unwanted additional tactile input, which could reduce sway variance (Pai & Patton, 1997). Second, the NeuroCom system has an increased cost compared to the pressure mat used in our system. Therefore, our system is available to a wider range of clinics and for at-home use. Last, the pressure mat used in our system is flexible, allowing it to be placed on different surfaces, characteristic of outdoor conditions.

FUTURE TRENDS

Primary trends in this area concern successful rehabilitation program completion. There are two main ways to achieve this: (1) further motivation by increasing the number and variety of games available to the user. One way to do this is through the development of an input device that will allow the COP signal from the pressure mat to play commercially available video games. In turn, this will keep the users motivated and interested and eliminate the need to develop new games; and (2) the progression of the exercise regime to an at-home program. The ability to customize our system to each subject's needs, the system's portability, and the embedded assessment file affords its use in monitored at-home programs. Additional work is required to communicate the assessment files to the clinician and allow the clinician to reconfigure the patient's game settings.

CONCLUSION

In this research, the need for and requirements of an effective rehabilitation regime were addressed. Of primary importance was motivating the patient to perform their exercises, using methods which are task-specific. Based on these requirements, an interactive video game based tool was developed to aid individuals with balance disorders and motivate them to complete their exercise program.

ACKNOWLEDGMENT

This work was funded by the Manitoba Health Research Council (MHRC) and Natural Sciences and Engineering Research Council (NSERC).

REFERENCES

Bach-y-Rita, P., Wood, S., Leder, R., Paredes, O., Bahr, D., Bach-y-Rita, E.W. et al. (2002). Computer-assisted motivating rehabilitation (CAMR) for institutional, home, and educational late stroke programs. *Topics in Stroke Rehabilitation, 8*(4), 1-10.

Betker, A.L., Szturm, T., Moussavi, Z., & Nett, C. (2006). Video game-based exercises for balance rehabilitation: Case study. *Archives of Physical Medicine and Rehabilitation, 87*(8), 1141-9.

Carr, J.H., & Shepherd, R.B. (1998). *Neurological rehabilitation: Optimizing motor performance.* Oxford: Butterworth-Heinemann Ltd.

Cheng, P.T., Wu, S.H., Liaw, M.Y., Wong, A.M.K., & Tang, F.T. (2001). Symmetrical body-weight distribution training in stroke patients and its effect on fall prevention. *Archives of Physical Medicine and Rehabilitation, 82,* 1650-1654.

Cogan, A., Madey, J., Kaufman, W., Holmlund, G., & Bach-y-Rita, P. (1977). Pong game as a rehabilitation device. In *Proceedings of the 4th Annual Conference on Systems and Devices for the Disabled* (pp. 187-188).

Cunningham, D., & Krishack, M. (1999) Virtual reality: A holistic approach to rehabilitation. *Studies in Health Technology and Informatics, 62,* 90-93.

De Weerdt, W.J.G., & Harrison, M.A. (1997). The use of biofeedback in physiotherapy. *Physiotherapy, 71,* 9-12.

Femery, V., Moretto, P., Hespel, J.M., & Lensel, G. (2001). Plantar pressure biofeedback device for foot unloading. In *Proceedings 5th Symposium Footwear Biomechanics*.

Gill, J., Allum, J.H.J., Carpenter, M.G., Ziolkowska, M.H., Adkin, A.L., Honegger, F. et al. (2001). Trunk sway measures of postural stability during clinical balance tests: Effects of age. *Journal of Gerontology Series A: Biological Sciences and Medical Sciences, 56A*(7), M438-447.

Glanz, M., Klawansky, S., & Chalmers, T. (1997). Biofeedback therapy in stroke rehabilitation: A review. *Journal of the Royal Society of Medicine, 90,* 33-39.

Harris, J.E., Eng, J.J., Marigold, D.S., Tokuno, C.D., & Louis, C.L. (2005). Relationship of balance and mobility to fall incidence in people with chronic stroke. *Physical Therapy, 85,* 150-158.

Huang, H., Wolf, S.L., & He, J. (2006). Recent developments in biofeedback for neuromotor rehabilitation. *Journal of NeuroEngineering Rehabilitation, 3*(11).

Jones, K.E., Wessberg, J., & Vallbo, A. (2001). Proprioceptive feedback is reduced during adaptation to a visuomotor transformation: Preliminary findings. *Neuroreport, 12*(18), 4029-4033.

Kwakkel, G. (2006). Impact of intensity of practice after stroke: Issues for consideration. *Disability and Rehabilitation, 28*(13-14), 823-830.

Lee, M.Y., Wong, M.K., & Tang, F.T. (1996). Using biofeedback for standing steadiness, weight-bearing training. *IEEE Engineering in Medicine and Biology, 15*(6), 112-116.

Marklund, I., & Klassbo, M. (2006). Effects of lower limb intensive mass practice in poststroke patients: Single-subject experimental design with long-term follow-up. *Clinical Rehabilitation, 20,* 568-576.

O'Connor, T.J., Cooper, R.A., Fitzgerald, S.G., Dvorznak, M.J., Boninger, M.L., VanSickle, D.P. et al. (2000). Evaluation of a manual wheelchair interface to computer games. *Neurorehabilitation Neural Repair, 14*(1), 21-31.

Pai, Y.C., & Patton, J. (1997). Center of mass velocity-position predictions for balance control. *Journal of Biomechanics, 30*(4), 347-354.

Peterka, R.J. (2002). Sensorimotor integration in human postural control. *Journal of Neurophysiology, 88,* 1097-1118.

Rougier, P. (2004). Optimising the visual feedback technique for improving upright stance maintenance by delaying its display: Behavioural effects on healthy adults. *Gait Posture, 19,* 154-163.

Sveistrup, H. (2004) Motor rehabilitation using virtual reality. *Journal of NeuroEngineering and Rehabilitation, 1*(1), 1-10.

Tsang, W.W., & Hui-Chan, C.W. (2004). Effects of exercise on joint sense and balance in elderly men: Tai Chi versus golf. *Medicine and Science in Sports and Exercise, 36*(4), 658-667.

Urhan, O., & Dincer, H. (2001). Gait trainer for children with spastic cerebral palsy. In *Proceedings 23rd IEEE Engineering in Medicine and Biology Conference, 2* (pp. 25-28).

Zambarbieri, D., Schmid, M., Magnaghi, M., Verni, G., Macellari, V., & Fadda, A. (1998). Biofeedback techniques for rehabilitation of the lower limb prosthetic subjects. In *Proceedings VIII Mediterranean Conference on Medical and Biological Engineering and Computing.*

KEY TERMS

Balance: The maintenance total body equilibrium.

Biofeedback: A biological signal which is recorded and presented to a person in real-time. In most cases, a visual or auditory feedback representation of the input signal intensity is used.

Center of Pressure: The directional sum of the vertical load exerted by the body.

Exercise: An activity performed in order to increase motor control.

Pressure Mat: A tool used to measure the load exerted by an object.

Rehabilitation: A therapy program used to improve motor control and balance.

Video Games: A game which is played on a personal computer.

Issues, Claims, and Concerns of Different Stakeholders of Healthcare Systems

Herwig Ostermann
University for Health Sciences, Medical Informatics and Technology, Austria

Bettina Staudinger
University for Health Sciences, Medical Informatics and Technology, Austria

Magdalena Thöni
University for Health Sciences, Medical Informatics and Technology, Austria

Roland Staudinger
University for Health Sciences, Medical Informatics and Technology, Austria

INTRODUCTION

Adopting a holistic sociotechnical perspective, healthcare systems do not merely exhibit complex structures and functionalities but are also affected by the differing expectations, claims, and concerns of the systems' stakeholders (Guba & Lincoln, 1989; Haux, Winter, & Ammenwerth, 2004). Furthermore, the issues addressed at healthcare systems are not limited to the concerns and requirements of health service providers, whose primary and most fundamental concerns in general terms represent the assurance of their own economic well-being and ability to proactively operate as well as the development of sustainable strategies in order to realize their own interests whatever they may be (Carsten, Hankeln, & Lohmann, 2004; Kappler, 1994). Furthermore, the objectives of other health systems stakeholders such as hospital operators and financiers as well as (health) politicians, which may well be in contradiction to the objectives of mere health service providers, have to be incorporated when systematically analyzing healthcare systems (Horev & Babad, 2005; Peltier, Kleimenhagen, & Neidu, 1996; Staudinger, 2004a).

If one now considers the capability of various stakeholders to implement their varying concerns, one may first assume that the political sphere in conjunction with hospital as well as healthcare operators and financiers has exclusive rights on organizing the system because of the common aptitude for developing and executing norms and directives in a way that allows the system's functionality to correspond to their own objectives. However, it has to be noted that in the case of health service providers delivering respective services within these system specifications, the subsequent adoption mechanism may have a reversing effect as compared to the original intentions and objectives of, for example, the legislators (Lim, Lee, & Taehun, 2005; Puxty, 1994).

In this context of potentially competing (and even conflicting) stakeholders' interests, the quality and availability of process data is of particular importance for all affected system participants. When considering the questions of system transparency and functionality, process data which are valid and, more importantly, able to be evaluated, provide the basis for judging whether certain system objectives are actually able to be implemented and have been implemented and whether deviations have taken place or not . However, as the process data is again acquired by different stakeholders themselves, conflicts may arise with regard to different perceived levels of process support and data management may well become an essential part of the conflict due to different data management objectives of the system's participants (Lim et al., 2005; Power, 1999).

BACKGROUND

The challenges involved with the potential conflict of competing issues, claims, and concerns of healthcare systems' stakeholders are currently being analyzed and observed on an international scale with a substantial amount of research focusing on the concept that participative models contribute to the solving of this conflict (Byrne & Sahay, 2007; Chaulagai et al., 2005). However, opposing this popular presumption of participative models are the health service providers' fundamental interests in being involved in the—particularly financial—health resource allocation processes in a manner which is conducive to allowing the greatest range of possible actions as well as reactive capability for themselves (Robinson, 2000; Robinson, 2004).

Playing an important role here is the fact that it is not only the health service providers, operators, financiers, and politicians striving for the realization of their concerns. Patients as well as tax payers also have an important say in this potential conflict and wish to incorporate their interests, too. Moreover, conflict-laden moments may also take place within different groups of stakeholders themselves. If one regards the general public, for instance, one may well realize competing multistakeholder interests when taking the stance of tax payers, where people may be less inclined to pay higher taxes on the one hand, and considering the interests of tax paying patients, who expect an adequately financed healthcare system, on the other (Staudinger, 2004a).

Hence, this resource-based approach to the implementation and evaluation of healthcare systems outlined above may serve as an appropriate tool for maintaining an overview of the development of the cost-service ratio (Webster, 2006; WHO, 2000). This approach, however, appears to be fairly inadequate because of two reasons: First, and more generally, healthcare systems do not necessarily adhere to market economy criteria. Second, the resource-based approach only considers non-monetary phenomena such as the local provision of healthcare structures when actual services are provided at this level and parameters such as service-quality ratio as well as the necessity for the provision of certain medical services can be taken into account (Berger, Honig, & Spatz, 2006). This problem of incorporating relevant phenomena is more intensified by the fact that, through the formation of (virtual) medical networks, the analysis of the individual health service providers from a strategic perspective is no longer sufficient (Waitzkin, Jasso-Aguilar, Landwehr, & Mountain, 2005).

Beyond pure cost considerations, analyses of the effectiveness of medical facilities and national healthcare systems therefore have to be carried out on the basis of input and output analyses, quality analyses , and associated process analyses. This in turn has to be done on the basis of process data which are (1) comparable and hence able to be analyzed on the one hand and (2) represent an integrative part of data management within the healthcare system on the other (Kaushal, Bates, Poon, Jha, & Blumenthal, 2005).

The challenges involved with the incorporation (and sometimes also unilateral resolution) of different stakeholders' issues, claims, and concerns into health systems shall be presented using the example of healthcare reforms in Austria.

ISSUES, CLAIMS, AND CONCERNS OF DIFFERENT STAKEHOLDERS OF HEALTHCARE SYSTEMS

Trade-Offs between Interdependent Stakeholders' Objectives in Austria

The Austrian healthcare systems can be first and foremost characterized by its decentralized and mostly federal nature in terms of decision-making as well as service-provision (Hofmarcher & Rack, 2006). As in other social insurance countries, virtually the entire national population is obliged to be health insured (save for a few exceptions such as soldiers, claimers of social benefits or prisoners) and people are bound by legislation to membership of a particular insurance association depending on the individual profession. The financing of hospitals and doctors, again, is in principle thought to be provided by social insurance associations too, even though the members' insurance premiums are not sufficient to finance the entire system. For this reason, federal government taxes are also incorporated in order to maintain a stable system (Hofmarcher & Rack, 2006).

In principle, each insured person is able to freely choose his or her hospital and physicians irrespective of whether these healthcare providers have contracts with the relevant insurance association or not. Financial compensation between different (in particular federal) insurance associations is only granted up to a certain extent, and in the past this has led to strong competition between the healthcare providers—and also between the insurance associations themselves (Hofmarcher & Rack, 2006).

Currently, the Austrian healthcare system is still lacking a reasonable amount of data exchange between healthcare providers as well as financiers, and neither central patient records nor process requirements or instruments for quality assurance exist (Pfeifer, 2001). Well aware of these shortcomings, the Austrian government had already decided upon a multilevel reform concept in 1997 in order to ensure the financing of healthcare and maintain cost-service transparency. Within this concept of reform, special attention was paid to the visionary objective of establishing a national network of communicating healthcare institutions for decentralized regional service provision. When devising the healthcare reform of 2005, this objective of networking healthcare providers of all different levels was determined in more precise terms and finally led to the concept of regional hospitals representing the vital nodal points of the intended networks (BMGF, 2006; Hofmarcher, 2006).

An important requirement in association with local (and also federal) integrative care structures featuring a hierarchical differentiation in terms of its functionalities is the challenge of reaching an agreement upon a common and minimally realizable set of objectives. Put in other words, by negatively limiting the task, healthcare systems planning consequently has to exclude contradictory interests of different stakeholders due to their inhibitory nature in order to guarantee a goal-orientated interdisciplinary service provision within a given care structure. (Bradley, Williams, Brownsell, & Levy, 2002; Ulrich & Wille, 1997).

As in other social insurance countries, the provision of healthcare services is regarded as an essentially public task in Austria, which is for the sphere of public healthcare subject to the principle of central healthcare planning and which is, at least in theory, free of market economy mechanisms. This implies that the organizational structure of public healthcare is based on the premise that a particular demand for healthcare provision has to be satisfied by a particular range of service offers and over- as well as under-capacities cannot or should not arise in the first place (Republik Österreich, 1990; Rosenbrock & Gerlinger, 2004).

However, this state-directed and essentially planned economic approach to organizing and maintaining the national public healthcare provision is prone to be challenged by a hardly controllable increase of healthcare institutions and providers of all levels which attempt to gain competitive advantages in the "health market" according to market economic principles. In the context of hospital and health systems planning, health politicians as well as civil servants of the health departments first have to consider that a multitude of adjustable and controllable elements of health systems is most commonly opposes the same multitude of freely operating health services providers. Moreover, the group of health system financiers is not congruent with the one of healthcare providers, which in turn are not identical to those politically responsible (Hofmarcher, 2006; Staudinger, 2004a).

In order to be able to correctly appraise and understand this basic framework of potential interests which interferes with healthcare planning, it appears rational to make the effort to re-examine what controls the actions and reactions of the system stakeholders on a more abstract level (Guba & Lincoln, 1989; Horev & Babad, 2005).

In Austrian healthcare, one can clearly identify a conflict of interest between hospitals providing health services on the one hand and the Federal Government, the *Länder,* and social insurance institutions representing interest groups involved in health politics and financing on the other (Staudinger, 2001). Here two different levels of conflict can be distinguished: service and cost (Table 1).

Regarding the service-level, the Federal Govern-

Table 1. Focuses of interest

Level of conflict	Federal Government, *Länder*, and Social Insurance Institutions	Hospitals
Service-level	<u>service optimization</u>	<u>service maximization</u>
	subsidiary principle	capacity principle
	appropriateness principle	availability principle
	minimal principle	optimal principle
Cost-level	<u>cost minimization</u>	<u>cost optimization</u>
	principle of structural diversification	principle of structural monopolization
	principle of statics	principle of dynamics

ment, the *Länder,* and social insurance institutions strive for the optimization of healthcare services in the sense of each individual patient receiving those healthcare services which are similarly optimal in terms of medical progress as well as business aspects. Hospitals and other healthcare providers are likely to take a different position in this respect as they are prone to incentives leading to the quantitative maximization of healthcare services provided. An example of this may be that in case of doubt, a patient will be taken in as an in-patient as long as there are beds free, even if out-patient treatment could well be sufficient under certain circumstances (Laimböck, 2001).

Moreover, the tenet of service optimization pursued by the Federal Government, the *Länder,* and social insurance institutions can be characterized by three phenomena, which are again opposed by three corresponding principles of healthcare providers seeking for service maximization:

1. Healthcare politics and financiers promote the subsidiarity principle, that is, the provision of healthcare should take place where relevant services can be provided at their lowest level possible. Hospitals, on the other hand, assume the capacity principle. This means that it should not only be the subsidiary institutions who provide care, but also all those who have the capacity to provide it. Illustrating an example, one could think of university hospitals which provide simple medical care when there is capacity for it, even if this care could be provided just as well in other facilities (as it is the case in Austria).
2. The Federal Government, the *Länder,* and social insurance institutions have an incentive to pursue the appropriateness principle. This means that the range of healthcare services provided to a certain patient is limited by the scale of services which are thought to be appropriate in order to achieve a certain medical outcome. Opposing this, hospitals apply the availability principle in the sense that it is not necessarily the reasonable and moderate range of services which is provided, but the services which are available (Ferguson & Smith, 1997).
3. In hospitals, it is the optimal principle which is applied rather than the minimal principle when it comes to actual service provision. This means that from a certain range of available healthcare services, all those will be chosen by hospital physicians which appear optimal for both the patient and the health services provider. Contrasting this optimal principle is the minimal principle which is endorsed by the Federal Government, the *Länder,* and social insurance institutions. Applying this minimal principle to the same variety of possible services available, only those services should be chosen whereby the desired outcome can just about be achieved with the least effort possible. Putting the optimal principle pursued by hospitals into practice, this may mean that if there is available capacity for an exact diagnostic examination using a computer tomography, it will be chosen instead of a conventional x-ray. Applied in a hospital laboratory context, all manner of analyses will be carried out as far as possible, and not just those deemed necessary for the case in hand.

Similar conflicts can be observed on the cost-level, too (Stepan, 1997): On the one hand, the Federal Government, the *Länder,* and social insurance institutions strive for cost minimization in order to achieve the lowest costs possible for the provision of a particular service. Hospitals, on the other hand, are rather inclined to the principle of cost optimization in a sense that for the maximal range and scope of services which the hospitals seek to provide costs should be optimized.

The principle of structural diversification is then endorsed by health politics and financiers in order to realize the overall objective of cost minimization. Put into practice, this principle requires that service-provision structures have to follow the according cost structures in a treatment chain. Hospitals, however, perceive this issue differently by pursuing the principle of structural monopolization which implicates the organization and management of institutional structures in such a manner that cost advantages are achieved not through diversification, but through the portfolio of services provided within a certain unit. This means that certain departments in hospitals should offer the most extensive care possible, and only within the departments themselves is it decided which unit actually provides which services. The conflicts triggered by the question of the differentiation of healthcare structures can be clearly observed in the current discussion about the spin-off of hospitals' out-patient departments in Austria, where redundant structures are still predominant with the same services provided to out-patients by hospital departments on the one hand and by physicians in private practice on the other (Fiedler, 1998).

Different objectives of hospitals as well as health politics and financiers can be moreover identified on the service-level when analyzing the principle of statics vs. the principle of dynamics. The former principle pursued by the Federal Government, the *Länder,* and social insurance institutions means that there should be as few changes as possible, as a low quota of change is commonly associated with a modest, if any, cost increase. In contrast to this, hospitals are rather inclined to apply the principle of dynamics. Pursuing this approach, healthcare providers attempt to promote and bring in as much innovation into the individual areas of healthcare as possible so that a maximization of services can be achieved through competitive advantages whilst achieving cost optimization at the same time.

Referring back to the original situation, a reciprocity of objectives can be identified between the service-level and cost-level in a way that the optimization of services and optimization of cost as well as the principles of service maximization and cost minimization are thought to behave in a reciprocal manner.

These conflicts of interests between healthcare systems operators and financiers on the one hand and hospitals on the other, and consequently between the different hospitals themselves and at the boundaries to practicing physicians' service portfolios of the private sector too, lead to additional costs as well as other financial burdens. However, it has to be taken into account that along with the expansion of costs over the last few decades, a substantial increase in the amount as well as quality of healthcare services provided has taken place, which may well be regarded as a predominantly positive occurrence (Becker & Beck, 1999).

If one now wants to assess this system of potentially competing objectives in association with a decentralized geographic service area according to the principles of integrated healthcare services provision, three systematic conditions have to be restored in order to resolve this conflict of interests (Staudinger, 2004b):

1. A sufficient convergence between the issues, claims, and objectives of all stakeholders of the healthcare system has to be created paying particular attention to higher priority objectives (Cuadras-Morató, Pinto-Prades, & Abellán-Perpiñán, 2001).
2. An up-to-date range of healthcare services provided has to be guaranteed for the entire population.
3. Healthcare structures which allow for a further sophistication of services have to be established.

The convergence of stakeholders' objectives was, for the most part, able to be achieved in Austria through the implementation of the 1997 structural reform and its accompanying measures for the in-patient sector (Laimböck, 2000). Taking into consideration the divergent objectives which were put forward by various interest groups when devising this reform, it appears to be valuable to take a closer look at the regulatory instruments with which the Federal Government and the *Länder* attempted to bring the objectives of hospitals as well as health system operators and financiers onto a common basis. As shown in Table 2, this attempt on establishing a convergence of objectives through legislation can indeed be described as a stipulated structural compliance from the hospitals towards the objectives of healthcare politics and financiers.

Considering the weighting of services provided and the evaluation of performance-oriented diagnosis-related case groups (PODRG) points it has to be noted, though, that these mechanisms do not necessarily have the desired control effect. Rather, healthcare providers, as seen in Austrian practices so far, still strive for service maximization which in turn leads to imbalances in the system.

Service optimization vs. service maximization can be attained through an obligatory national health plan based either on service provision planning or on the more traditional planning of locations and specialties offered (Hofmarcher, 2006). By the means of promoting the expansion of the extramural sector, it is intended to strengthen the subsidiarity principle. Hospital funding via PODRG points has been introduced in order to transparently record and demonstrate the appropriateness of medical services provided, and the conducted weighting of the services and LDF points is thought to support the implementation of the minimal principle when it comes to the actual service provision (Zweifel, 1997).

The most considerable cut on the cost-level represents the introduction of capped hospital budgets, as hospitals are now actively encouraged to shift their strategic focus from cost optimization to cost minimization. Moreover, the measures taken here encompass the

Table 2. Means of convergence

Federal Government, *Länder*, and Social Insurance Institutions	Hospitals	**Means of Convergence**
service optimization	service maximization	**obligatory hospitals and major equipment plan**
subsidiary principle	capacity principle	**expansion of extramural service provision**
appropriateness principle	availability principle	**PODRG points***
minimal principle	optimal principle through LDF points	**weighting of services provided and PODRG points**
cost minimization	cost optimization	**capping of budgets**
principle of structural diversification	principle of structural monopolization	**revenue-oriented expenditure policy**
principle of statics	principle of dynamics	**additional specific funding of university teaching hospitals**
* PODRG: Performance-orientated diagnosis-related case groups. The number of points assigned to a PODRG determines the PODRG flat rate (Hormarcher & Rack, 2006).		

implementation of revenue-oriented expenditure policy (administered by so-called *Länder*-funds) as well as the introduction of additional controlled financing for the university teaching hospitals in order to curb the system immanent principle of dynamics.

Functionality Requirements of an HIS Caused by Trade-Offs between Stakeholders' Objectives

The political requirements intended to serve as means of convergence in the conflict of interest outlined above (see Table 2) force healthcare providers and hospitals in particular to redefine their own information management, hospital information systems (HIS) and logistics in order to prevent redundancies as far as possible (Kaushal et al., 2005). System adapted behavior does not allow for in-patients remaining too long in hospital due to inadequate process support of in-house information management or a diagnostic laboratory working too slowly in order to merely attain an unambiguous diagnosis.

Consequently, the standardization of medical services and in particular processes has become more significant over the last decade and this in turn has resulted in two further consequences: (1) Whilst information management played a rather insignificant role in Austrian hospitals up to the implementation of the 1997 healthcare reform due to the objective of enhanced transparency and comparability of institutions being of no systematic importance, HIS today play an essential role in the (re-)engineering and management of hospital processes. (2) The integration of extramural facilities represents another important consequence, as strategic planning within hospital management appears to be indispensable in order to achieve continuity and smooth patient pathways within the treatment chain (Staudinger, 2004b).

The primary requirement of an HIS, caused by the consequences of system adapted behavior outlined above, is the capability to serve as a nodal point for different healthcare providers in order to ensure their integration into healthcare networks (Montreuil & Garon, 2005). Here, one has to consider that the different healthcare providers in question have to be integrated along treatment chains as a whole (Kaushal et al., 2005), but may well feature different typologies and profiles, which again plays a vital role when establishing networks and even when just expanding communication structures.

The bases of each HIS are patient and treatment data in context with process management and control, which again have to reflect the different typologies of healthcare providers in more than just one sense:

1. First the question of the validity of the data has to be considered in order to figure out whether the diagnostic findings of one particular healthcare provider in the context of a treatment process will be able to be taken on by the subsequent treatment units without further examination. Criteria of common data security have to be de-

cided upon and implemented in order to prevent time-consuming multiple examinations, which lead to an unnecessary increase in cost. Requirements of plausibility examinations are therefore of particular importance when implementing and operating integrative HIS.

2. The question of the availability of data and data access also needs to be addressed in context with different types of healthcare providers participating in a particular treatment chain. According to the corresponding norms of the Austrian Data Protection Act of 2000, within the context of an integrated chain of treatment only those data sets may be made available to the service-providing physician or health professional which are sufficient to carry out a particular treatment. All other patient data are subject to data protection (Schwamberger, 1999).
3. This fact consequently leads to the next challenge of data management addressing the question of which minimum data sets have to be available to all participating facilities in order to cover both the legal and medical requirements. Simultaneously, it has to be considered that the financiers and political system operators both demand feasible decision transparency within the treatment process also in order to able to verify whether financial and political systems requirements are translated into practice by healthcare providers.

Summing up, this means that within an HIS, a particularly defined and differentiated authorization system has to be established for data access and at the same time a consensus about a minimum data setting encompassing all healthcare providers as well as processes has to be achieved or maintained. Furthermore, additional authorizations for the entire treatment chain have to be realized for the specially developed processes which reflect predefined treatment paths (Haak et al., 2002).

Another vital requirement of an HIS represents the functionality of integrative resource planning and control in a way that medical healthcare provision can be planned throughout the entire treatment chain and that individual services can be incorporated into the total planning irrespective of the actual provider (Bradley et al., 2002). The importance of this capacity is in particular due to the fact that any HIS has to be combinable with the structural planning of geographic service areas as well as with the accounting system demanded by the finance mechanism.

Apart from these institutional requirements, the patient also has to be incorporated into the entire system too, as it is the patient who has to give consent to all services provided throughout the treatment chain and hence is in need of information about the whole devised treatment process. Moreover, the patient taking the role of an independent and responsible partner in the system still has the choice of which healthcare provider he or she charges a particular healthcare service with, irrespective of a predefined treatment chain. If the patient now decides not to stick to the predefined treatment path, data management and HIS then face the challenge of allowing for customized processes to be included (Leonard, 2004).

The basic challenge outlined above, namely the solution of the structural and institutional conflict between different stakeholders, can be addressed by the implementation of integrative HIS, however, to a rather limited extent. On the one hand, medical processes have to be supported and structural requirements have to be implemented as specified above. On the other, it is in the objective of healthcare providers—even if they are part of regional service networks—to define isolated positions and aims and make them the basis of their future operations.

An efficient HIS now shows the aptitude to indirectly mitigate the conflict within the context of the described structural conditions as it substantially contributes to the cost-effectiveness of the whole range of processes and makes a wider database available to stakeholders, which enables quicker decisions and more accurate planning. Moreover, this common database bears the advantage for all system participants of a comparability of different means of healthcare service provision, which in turn allows for a clearer definition of the scope of operations and reforms and therefore also for the increase of one's own efficiency.

FUTURE TRENDS

The different issues, claims, and objectives of healthcare system stakeholders such as patients, healthcare providers, and financiers as well as political decision-makers are likely to prevail in the future and can be expected to still have an impact, even on the level of individual medical decision-making. The current trend of merging healthcare providers into networks on the basis of treatment paths and geographic service areas in the context with a uniform financial system is able to mitigate this problematic of conflict of interests, but not able to ultimately solve it. As far as data management is concerned, one therefore has to take into consideration that different system participants may also have different requirements of an HIS and of integrated data management, and that at the same time, IT-management finds itself right in the middle of this conflict (Staudinger, 2004b).

One may therefore well assume that due to changed structures of healthcare service provision, the requirements on data management will increase in various ways: on the one hand, technical problems such as the ability of different HIS to correspond and the question of data quality and process support have to be addressed. On the other hand, the systemic specifications of data management are also likely to increase, especially when these data form the basis for healthcare system planning or medical controlling. It is here in particular that there is the opportunity for information technology to support the change processes in healthcare institutions and facilities so that individual interests can be suppressed or modified to conform to higher level objectives. This in particular concerns the cost-cutting and service increasing tasks such as knowledge management, quality assurance in the medical sector, the integration of patients into the information system, and the combining of medical healthcare providers with commercial systems as well as in the area of planning and resource management (Hain, 2002; Power, 1999; Smirnov, Pashkin, Chilov, Leveshova, & Krizhanivsky, 2005). In all of these areas, a win-win situation can be brought about through information management which corresponds to the general objectives of all system stakeholders.

CONCLUSION

Healthcare systems seen on an international level are complex structures which are influenced by the differing objectives of system stakeholders. In contrast to pure market economy systems whereby the balance of interests is maintained (in theory) through market mechanisms, healthcare systems define themselves as systems which are planned and controlled from the supply-side even though being ultimately demand-induced. In these systems legislators and financiers seek to gain influence on decision processes, system profiles and system suppliers by the active management of the general regulatory, political and financial framework.

Using the example of Austrian healthcare reforms of the last decade, different operating principles and objectives of healthcare system stakeholders were portrayed in detail. Through the implementation of a performance-orientated financing system as well as through elements of central planning and specific monetary incentives, the Federal Government, the *Länder,* and social insurance institutions attempt to accomplish their objectives of subsidiary healthcare provision, cost orientation, and transparent decision making. On the other hand, healthcare providers and, in particular, hospitals, attempt to enforce their own objectives, which diverge with the above mentioned interests of health politics and financiers. The interest of healthcare providers encompass the provision of a maximal range of services for patients, the retention of a service-induced profile and maximal integration of medical advances in treatment paths whilst incorporating maximum facility independence. Through the mechanisms implemented by the healthcare reforms, healthcare providers nowadays are forced into stronger networking, transparency, and standardization, without having to completely give up their own objectives.

The new mechanisms of healthcare finance as well as planning have then brought new challenges for IT management, particularly in hospitals. A stronger rate of HIS implementation could be observed in first stage and consequently the networking of different healthcare providers became a main challenge for data management. Of particular importance here was the support in the area of medical process standardization as well as the question of communication ability between the facilities. In association with these requirements was the question of data protection and quality assurance of data as well as the definition of a minimum data set.

Only once these requirements have been met in a sustainable manner will an integrative information system be available to the healthcare providers as well as politics and financiers, which will at least partially be able to contribute to the solution of the conflict of interests through the bringing about of win-win situations, particularly in the areas of resource management, cost transparency, and process security.

REFERENCES

Becker, A., & Beck, U. (1999). Gegenstrategie zum Parteivergleich: Wie bereitet sich ein Krankenhaus auf die Auseinandersetzug mit den Krankenkassen vor? *Das Krankenhaus, 9,* 583-589.

Berger, M., Honig, P., & Spatz, I. (2006). Medicare and cost-effectiveness analysis. *New England Journal of Medicine, 354,* 208.

BMGF (Ed.). (2006). *Österreichischer Strukturplan Gesundheit 2006.* Vienna, Austria: Federal Ministry for Health and Women.

Bradley, D. A., Williams, G., Brownsell, S. J., & Levy, S. (2002). Community alarms to telecare: The need for a systems strategy for integrated telehealth provision. *Technology and Disability,* 14, 63-74.

Bynre, E., & Sahay, S. (2007). Participatory design for social development: A South African case study on community-based health information systems. *Information Technology for Development, 13,* 71-94.

Carsten, A., Hankeln, C., & Lohmann, R. (2004). Entwicklung und Implementierung von Strategien im Krankenhaus mit Hilfe einer Balanced Scorecard. *Journal für Anästhesie und Intensivbehandlung, 13*(1), 98-104.

Chaulagai, C. N., et al. (2005). Design and implementation of a health management information system in Malawi: Issues, innovation and results. *Health Policy and Planning, 20,* 375-384.

Cuadras-Morató, X., Pinto-Prades, J.-L., & Abellán-Perpiñán, J.-M. (2001). Equity considerations in healthcare: The relevance of claims. *Health Economics*, 10, 187-205.

Ferguson, B., & Smith, P.-C. (1997). Healthcare financing in the United Kingdom: National health service. In A. Stepan (Ed.), *Finanzierungssysteme im Gesundheitswesen: Ein internationaler Vergleich* (pp. 89-110). Vienna, Austria: Manz.

Fiedler, E. (1998). Kooperation zwischen stationären und extramuralen Betreuungseinheiten. In A. Grundböck, P. Nowak & J. M. Pelikan (Eds.), *Neue Herausforderungen für Krankenhäuser: Qualität durch Gesundheitsförderung – Gesundheitsförderung mit Qualität* (pp. 277-283). Vienna, Austria: Facultas Universitätsverlag.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park: Sage Publications.

Haak, M. et al. (2002). Networking in shared care: First steps towards a shared electronic patient record for cancer patients. *Methods of Information in Medicine, 41,* 419-425.

Haux, R., Winter, A., Ammenwerth, E., & Brigl, B. (2004). *Strategic information management in hospitals: An introduction to hospital information systems*. New York: Springer.

Hain, T. (2002). Improving the quality of health information: The contribution of C-H-i-Q. *Health Expectations, 5,* 270-273.

Hofmarcher, M. M., & Rack, H.-M. (2006). Austria: Health system review. *Health Systems in Transition, 8*(3), 1-247.

Horev, T., & Babad, Y. M. (2005). Healthcare reform implementation: Stakeholders and their roles: The Isreali experience. *Health Policy, 71,* 1-21.

Kappler, E. (1994). Das Krankenhaus der Zukunft aus ökonomischer Sicht. In E. Kappler & B. H. Mühlbauer (Eds.), *Kommunikationstechnik im Krankenhaus: Meinungen, Perspektiven, Auswirkungen* (pp. 10-16). Darmstadt, Germany: GIT Verlag.

Kaushal, R., Bates, D. W., Poon, E. G., Jha, A. K., & Blumenthal, D. (2005). Functional gaps in attaining a national health information network: What will it take to get there in five years? *Health Affairs, 24,* 1281-1289.

Laimböck, M. (2000). *Die soziale Krankenversorgung zwischen Staat, Monopol und Wettbewerb: Vor- und Nachteile des Wettbewerbs im Gesundheitswesen*. Hall/Tyrol: Berenkamp.

Laimböck, M. (2001). Die Verknüpfung und Kooperatione zwischen ambulanter und stationärer Krankenversorgung in Österreich. In C. Kaloczy, S. Kindermann & R. Staudinger (Eds.), *Krankenhausorganisation und -finanzierung* (pp. 173-188). Hall/Tyrol: Berenkamp.

Leonard, K. J. (2004). The role of patients in designing health information systems: The case of applying simulation techniques to design an electronic patient record (EPR) interface. *Healthcare Management Science, 7,* 275-284.

Lim, G., Lee, H., & Taehun, K. (2005). Formulating business strategies from a stakeholder's perspective: Korean healthcare IT business cases. *International Journal of Information Technology & Decision Making, 4,* 541-566.

Montreuil, B., & Garon, R. (2005). Toward patient centric and distributed healthcare delivery networks. *Clinical and Investigative Medicine, 28,* 351-352.

Peltier, J. W., Kleimenhagen, A. K., & Neidu, G. M. (1996). Integrating multiple publics into the strategic plan: The best plans can be derailed without comprehensive up-front research. *Journal of Healthcare Marketing, 16*(1), 30-37.

Pfeiffer, K.-P. (2001). Fünf Jahre Erfahrung mit der Leistungsorientierten Krankenhausfinanzierung (LKF) in Österreich. In P. Eichhorn, H.-J. Seelos & J.-M. Schulenburg (Eds.), *Krankenhaus-report 2001* (pp. 181-189). Stuttgart, Germany: Schattauer.

Power, M. (1999). An approach to policy analysis and development of medical informatics. *Methods of Information in Medicine, 38,* 260-264.

Puxty, A. G. (1993). *The social and organizational context of managerial accounting*. London: Academic Press.

Republik Österreich (Ed.). (1990). *Neuordnung der Kompetenzverteilung in Österreich: Grundlagen und Ergebnisse der Arbeitsgruppe für Fragen der Neuordnung der Kompetenzverteilung*. Vienna, Austria: Verlag der Republik Österreich.

Robinson, J. C. (2000). Deregulation and regulatory backlash in healthcare. *California Management Review, 43*(1), 13-33.

Robinson, J. C. (2004). Consolidation and the transformation of competition in health insurance. *Health Affairs, 23*(6), 11-24.

Rosenbrock, R., & Gerlinger, T. (2004). *Gesundheitspolitik: Eine systematische Einführung*. Bern: Verlag Hans Huber.

Schwamberger, H. (1999). Einige gesundheitsrechtlich relevante Aspekte des Datenschutzgesetzes 2000. *Recht der Medizin, 131*.

Smirnov, A., Pashkin, M., Chilov, N., Leveshova, T., & Krizhanivsky, A. (2005). Knowledge logistics as an intelligent services for healthcare. *Methods of Information in Medicine, 44,* 262-264.

Staudinger, R. (2001). Aspekte der Krankenanstalten- und Gesundheitssystemplanung. In C. Kaloczy, S. Kindermann & R. Staudinger (Eds.), *Krankenhausorganisation und -finanzierung* (pp. 137-148). Hall/Tyrol: Berenkamp.

Staudinger, R. (2004a). Gesundheitswesen und die Rolle der Politik. In E. Hauke & E. Holzer (Eds.), Gesundheitswesen: vom Heute ins Morgen (pp. 91-99). Vienna, Austria: WUV Universitätsverlag.

Staudinger, R. (2004b). Strukturvoraussetzungen für ein integratives und vernetztes Gesundheitssystem. *Informatik, Biometrie und Epidemiologie in Medizin und Biologie, 35,* 172-174.

Stepan, A (1997). Bemerkungen zu den Fallstudien. In A. Stepan (Ed.), Finanzierungssysteme im Gesundheitswesen: Ein internationaler Vergleich (pp. 33-76). Vienna, Austria: Manz.

Ulrich, V., & Wille E. (1997). Deutschland. In A. Stepan (Ed.), Finanzierungssysteme im Gesundheitswesen: Ein internationaler Vergleich (pp. 33-76). Vienna, Austria: Manz.

Waitzkin, H., Jasso-Aguilar, R., Landwehr, A., & Mountain, C. (2005). Global trade, public health, and health services: Stakeholders' construction of the key issues. *Social Science & Medicine, 61,* 893-906.

Webster, P. (2006). Controlling the costs of US healthcare. *The Lancet, 367,* 639-640.

WHO. (2000). *The world health report: Health systems: Improving performance*. Geneva, Switzerland: World Health Organization.

Zweifel, P. (1997). Schweiz. In A. Stepan (Ed.), *Finanzierungssysteme im Gesundheitswesen: Ein internationaler Vergleich* (pp. 77-88). Vienna, Austria: Manz.

KEY TERMS

Healthcare Financier: Persons or organizations funding the provision of healthcare services.

Healthcare Network: A (regional) group of healthcare providers offering differentiated ranges of services in a coordinated manner.

Healthcare Provider: Persons or organizations providing health services to third persons and, in particular, patients.

Healthcare System: The system assigned with the preservation of mental and physical health through services offered by physicians and other health professionals.

Medical Treatment Chain: The organized sequence of healthcare services for a certain indication.

Patient Pathway: The sum and sequence of healthcare services offered to a certain patient by various healthcare providers in order to achieve a certain medical outcome.

Stakeholder: A person or organization having a genuine concern or interest in a certain system or entity.

I

IT-Based Virtual Medical Centres and Structures

Bettina Staudinger
University for Health Sciences, Medical Informatics and Technology, Austria

Herwig Ostermann
University for Health Sciences, Medical Informatics and Technology, Austria

Roland Staudinger
University for Health Sciences, Medical Informatics and Technology, Austria

INTRODUCTION

Today, medical infrastructures are subject to organisational change the world over. The reasons for this are manifold. On the one hand, it can be observed through scientific innovation and gaining of knowledge that a more in-depth specialisation is taking place, which means that medical healthcare providers are able to offer in-depth knowledge in narrowing fields. On the other hand, through increased process orientation of treatment pathways for patients, the necessity for superior organisational principles has been established. This effects that organisational cooperation models have to be found, which integrate singular specialised institutions into an organisational whole, which then employ integrative processes, information, and quality requirements.

Such organisational structures can be set up in real structures by the spatial accumulation of scattered service units, so that centralised medical centres establish. Alternatively, organisational and technical integration may substitute the physical integration. The individual medical service units stay dislocated, but appear as a virtual medical centre or as specialised medical networks having a clearly defined profile.

The management of the information technology in a virtual medical service centre is subject to different requirements than the IT-management of a hospital resembling more a closed shop.

Building a virtual centre calls for performance of an open shop principle, because the entire treatment chain cannot be mapped within one single institution, but requires integrated cooperation in order to manage a patient's clinical pathway.

Not only the spatial displacement, but also the unavoidable higher process orientation within a virtual cooperation deserves particular consideration.

Additionally, the information management is challenged by the fact that the provision of relevant information in standardised form is an indispensable element of a virtual centre. In this context, the question about potential structural assembling, and organisational principles and elements of virtual medical service centres has to be answered in order to conclude on the basic requirements of data management and the appropriate solution approaches.

This shall be presented partly using the example of the virtual oncological medical centre in Tyrol.

BACKGROUND

The question about the role of medical informatics in modified organisational structures has been frequently posed in the context of medical service supply (Power, 1999). In particular, an integrative IT-policy has considerably increased in importance through the standardisation of patient pathways, introduction of evidence-based medicine, standards, guidelines, and directives, and therefore, the inevitable necessary cooperation of healthcare providers. Especially in the earlier stages, exemplary models were introduced which were supposed to make the changed requirements more manageable (Ölvingson, Hallberg, Timpka, & Lindqvist, 2002). This, however, does not only affect the cooperation between healthcare providers, but also the technical-based integration of patient data, as well as the extensive bonding of other persons and units involved with the treatment chain. Particularly opened by the evolution of telemedicine, concepts are currently being developed in this area (Bradley, Williams, Brownsell, & Levi, 2002), which integrate technical possibilities such as virtual reality into medical treatment processes (Burdea, 2003).

The first and pivotal approaches of information technology in medical networks primarily concentrate on the field of electronic patient record. Correlating to this, scientifically evaluated projects and models introduced criteria which have to be considered by the IT-management in medical networks (Van den Haak, Mludek, Wolff, Bützlebruck, Oetzel, & Zierhut, 2002). It can therefore be adhered to, that through this structural development of the various sectors of the health system which are geared towards integrated care, medical informatics becomes more complex, too. Increased dialogue structures, principles of the standardisation of terminology, data exchange, and the comparability of data, and the resulting necessary transparency of medical healthcare providers have become subject matters of discussion and research (Coddington, 1997; Francis & Hart, 1998).

Beside the altered basic conditions, it can be observed that the aspired construct of "patient oriented health networks" actually take shape by the affiliation of medical disciplines towards patient oriented service centres, and hence in practice, forces a complex IT-management (Montreuil & Garon, 2005).

Linked to this is not only the question of the requirements and the practical implementation of such systems (Day & Norris, 2006), but also the investigation about barriers and restraints of a successful realisation of new concepts (Cashen, Dykes, & Gerber, 2004). At least the differing intentions and interests of single system partners (e.g., administration) also play an important role (Hassan, 2005), as do the requirements of health politics (Kaushal, Bates, Poon, Ashish, & Blumenthal, 2005). Dealing with the subject of virtual medical centres, science particularly has to focus on standardising the processes in the networks, and also on the quality of data (Hain, 2002; Stoop & Berg, 2003).

IT-BASED VIRTUAL MEDICAL CENTRES AND STRUCTURES

The Model of the Virtual Medical Centre

Based on the newer findings about organisational concepts and structures in medicine, Tyrolean (Austria) health politics initiated a project in 2000, whereby it was attempted to integrate oncological healthcare providers in the province of Tyrol into a virtual network. The network was supposed to be implemented on the basis of uniform medical standards coupled in clearly defined treatment paths. The participation in the virtual medical centre was voluntary, but all participants had to comply with the commonly derived directives. Simultaneously, the basic organisational principles of the virtual oncology centre were developed in order to be able to work out the requirements for the subsystems (e.g., the data management).

The basic principle of the virtual medical centre in Tyrol can be described through structural tripartism into strategical, tactical, and operative level. On a processual level, this principle can be best compared with a cybernetic model, whereby the learning ability of the system is centrally controlled and ensured by the interaction of all system participants.

A medical board is operating on the strategic central level, its main task being that of serving as a steering board for the entire medical centre. Its particular function is to validate directives and processes, as well as to adjust system parameters in the area of quality assurance. Complimentary, those carriers who additionally to medical competency have the key competencies for steering the entire system are part of the board (e.g., clinic management executives).

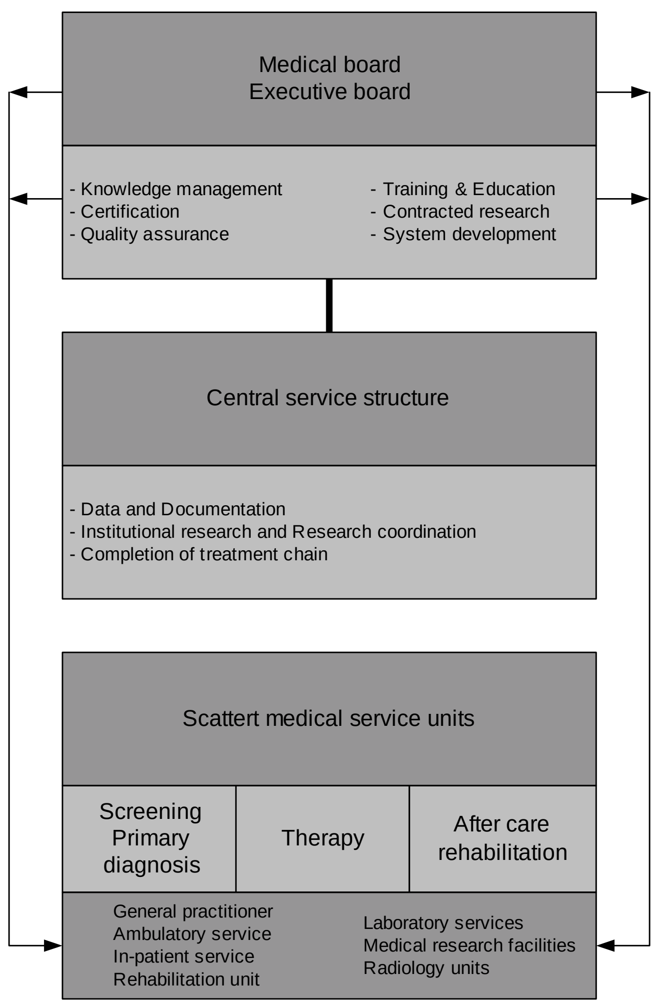
As contract research and pursuit of research projects are related on the board level, an "Open Shop" solution for communication instruments has to be provided (Ohly, 1995). For a consequence, communication and interaction structures have to be connected beyond the virtual medical centre. Pipelines to supraregionally and internationally active healthcare and research centres are provided, as well as structural ties to industrial partners involved with research.

As a tactical element, the central service structure particularly serves data and patient management, and has to include all documentation systems which are of superior importance. Irrespective of this, a regionally spanning consensus still has to be reached about documentation criteria—of patient files for example (Leiner, Gaus & Haux, 1995). This appears feasible if well-described and tested models are employed (Yamasaki & Satomura, 2001). Furthermore, own research projects as well as institutional research is steered and coordinated on the central service level. Treatment structures are only to be maintained up to a certain degree as gaps in the treatment chain within the framework of decentralised efficiency structures have appeared, and cannot be qualitatively and quantitatively closed by decentralised healthcare providers.

The decentralised medical service units concern patient-related healthcare providers who have been certified by the medical board on the basis of the derived standards and quality directives, as well as the resolved treatment chain. The healthcare providers are therefore responsible for screening right through to aftercare in the areas of standard patient medicare and the provision of research related medicare (Ruggiero, Giacomini, Sacile, Sardanelli, Nicolo, & Lombardo, 1995). These decentralised service structures are represented by health facilities or sectors of health facilities according to which fields were certified in the cooperation agreement.

The participation of the healthcare providers as part of a virtual medical centre is on a voluntary basis and is decided by mutual rights and obligations, or rather, by agreed behaviour modalities of the system participants amongst themselves. The area of mutual obligations primarily refers to the willingness to accept and employ standards and directives—which were defined on the part of the board in terms of "good clinical practice" and were subject to international evaluation—as well as the willingness to adapt clinical, therapeutic, and other processes and procedures according to these directives and guidelines (Medical Committee, 1999). This means that certification by the board has to take place when a participant enters the entire system, whereby the objectives of structural networking and system compatibility in particular have to apply, as well as the aspects of training standards and quality assurance. In

Figure 1. Model of a virtual medical competence centre



addition, through certification, it has to be ensured that the processes and procedures, as well as documentation and treatment plans, are adapted to the ongoing and continually changing basic conditions and directives, and therefore correspond to the processual element of medical development. This means that certification cannot be one singular act which is coupled to participant entry into the system. It is rather a periodically reoccurring procedure which evaluates the content and quality of the changed processes.

As the healthcare providers on this operative level face each other in an interactive and quality assured network, material and formal conformity and transparency of processes and service profiles are ensured (Göbel & Pfeiffer, 2001), and patients can be led through the entire treatment process without having to question the structure quality of a single network participant (Schoop & Wastell, 2001). Out of this basis, the mutual rights of the participants are able to be derived. These mutual rights consist in particular of gaining shortest possible and interpreted access to new medical findings. In conjunction with standardised training programmes and a superior information management, this leads to system participants gaining both medical and forensic confidence. Additionally, competencies and functions which have to be purchased from third parties (e.g., laboratory reports) by single system partners may be integrated simply standardised into their own operational structuring respectively into their processes. A major advantage on this level is the possibility to gain an overall (anonymous) data overview, and to have valid and broad data material available in order to put research cooperation or internal research projects onto a basis, and to find the current status of research prepared in its final form (Schnabel, 1996).

Cooperation Directions Within the Network

Not all descriptions of a particular structure or process have the same degree of validity. It is therefore important to make the distinction that the degree of commitment of a directive is clearly higher than the degree of commitment of a guideline which, as the name suggests, should provide guidance in a particular section of the process but with less commitment than is the case with a directive. Both directives and guidelines move within defined medical standards, whereby this term is not without controversy in connection with the history of standards in literature. In particular, newer works point to the problem that within defined standards (e.g., in the branch of clinical pharmacology), industrial interests can be found (Tenery, 2000; Wazana, 2000). This question is of importance, particularly for the medical director, as he is primarily responsible for introducing standards, guidelines, and directives, according to Austrian hospital law (Staudinger & Mair, 2001).

Important for describing a virtual medical centre is the semantic discrepancy of the terms "virtuality" and "centre." These prima-facie of alternative concepts can only then be combined if the centralisation term is not seen in a spatial sense, but rather in its substantial core, and the spatial dimension is reduced to a roughly fenced territory.

This network approach—aggregating decentralized healthcare providers to a nonmaterial entirety on an operational level, where, in a spatial sense, central strategic, and tactical elements are effective—corresponds to the cybernetic model approach, as long as self-learning elements from interactive processes are integrated.

Within such centres, the subsidiary principle applies to the healthcare providers. Simply put, each of the respective healthcare providers, who just is qualitatively and quantitatively in the position to render a particular service, actually renders this service. It can be assumed that an optimal degree of decentralisation can be achieved with a system-immanent service distribution (Sacile, 1995), but rather in context with a qualitative service hierarchy than a structural one.

Structural and Functional Requirements for Data Management

Against the given background, it seems necessary to develop or engage communication tools which ensure process and directive orientation, as well as the interaction ability of healthcare providers in an entire system (Geißler & Rump, 1999). Linking elements of process-orientated resource planning and optimization items of networked structures additionally arises necessary from the fact that self-learning systems with a high degree of networking and interaction are deeply exposed to a considerable pressure, regarding qualitative and quantitative outcome, as well as cost-effectiveness.

Apriori, it can be assumed that data management is carried out on central parameterizable systems. A medical clinical information system (CIS) serves as a basic platform (Steffens, 1995) in which all communication steps necessary for patient care can be carried out. Various types of documents, such as patient-related messages, incoming and outgoing reports for diagnostic results, or other confidential data are transported. An entire hospital information system (HIS) is eked by additional functions which cover all areas not directly consisting of medical documentation, but serve as a basis in all branches of care delivery. Namely order management, patient management, and administration processes serve as examples here (Kaloczy, 2001). A key element of the hospital information system is the compilation of diagnoses and service delivery (usually from standard catalogues such as ICD-9 and ICD-10). As a rule, the catalogues aim at several objectives of patient documentation. Beside the feasibility of diagnoses and medical treatment also, relevant balancing aspects are included (Pernice, Doare, & Rienhoff, 1995).

These IT tools have to be implemented according to a specified profile that is aligned with strategic parameters, in order to be able to ensure the corresponding planning and optimisation basis of the entire system, and subsequently, to prevent undesired maximisation activities of individual system participants.

This integrative point of view of communication technology in health systems, taking virtual medical centres into particular consideration, impedes an over centralisation of data and process management or process and resource planning, rather it advances regional integration and consistent regional structuring of health facilities. Thereby, both individual tracts of the system, as well as the entire system have to remain able to communicate and interact with other regional health structures. This is of particular importance where the patient joins different parallel health systems, as patients might have need for health treatment beside the oncological treatment.

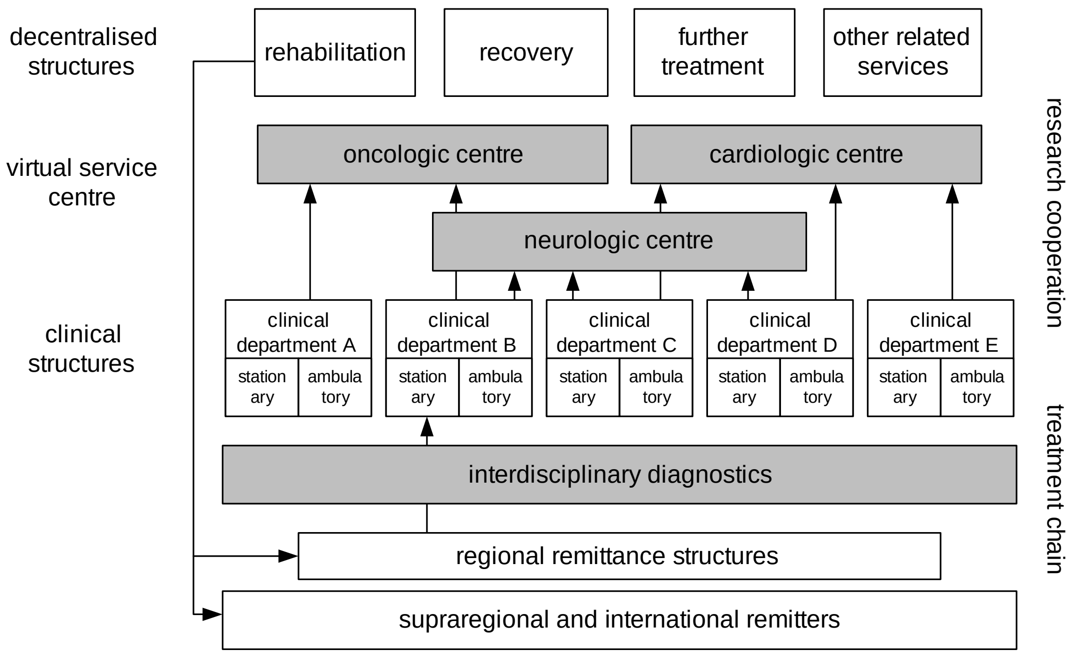
The fact that communication systems in context with virtual medical centres quasi take on the function of a transmitter is also of particular importance. Consequently, standardised communication within the predefined processes shall be able take place at any appearance. A transfer of data and results, as well as planning processes (Bachert & Classen, 1995) from one system participant to another, as well as within the whole system on a dialogue-orientated basis is therefore assured. Data redundancies and subsequent error vulnerability in data structures may be minimised (Gustafson, Hawkins, Boberg, Pingree, Serlin, & Graziano, 2001).

The reorganisation of a clinical information system and regional interaction system is chained with additional considerable cost and workload (Becker & Arenson, 1995), which can only then be justified when it opposites a benefit. This striven benefit in turn has to legitimate that a change in a running and functioning system leads to irritations within the operational structuring. In particular, each facility which shows a partial monopoly in oncological care is especially affected by this paradigm shift. On evaluating the new information system according to strategic objectives, the focus may not only be on technological solution, but rather the superior aim of organisational development which is compatible with strategic bases and development potential (Haux, 1995) must be drawn into account.

The self-conception and right to exist of single medical service structures are defined—along side other important factors—by the catchment area and prominence within the treatment chain. The catchment area neither has fixed size nor is it legally determined—leaving the basic political contract for primary healthcare out of equation. For healthcare providers in association with a virtual oncological medical centre, this means that their external communication structures have to be clearly defined to patients and remitters and have to assume a activity standard which ensures stronger quantitative and qualitative integration into the area of competence. Viewing the inner structure of the communication system operation and communication structures not only have to satisfy the current clinical requirements, but also must contain a cost and service orientated component (Kyriacou, Davatzikos, Zinreich, & Bryan, 2001).

The integration of the remitter (e.g., general practitioner, practicing specialist, primary care centres) into the clinical information system aims at a change from concerned people into partners. The question whether an organization member is part or not of a legal unit is of superior relevance, if integrative structures are set up. Rather the degree of service and information quality, which is achieved by the system change, comes to the foreground and may be an essential competitive advantage beside the clinical performance and efficiency. This assumes however, that the quality of communication and information within the hospitals or between the hospitals and the decentralized healthcare facilities achieves a level that enables to implement patient-orientated operating procedures. Whilst commercial

Figure 2. IT conceptions using Tyrol as an example



systems, particularly in the hospital branch, show a high and extensive functional state of development, medical systems are of a minor evolution standard (Clayton, 2001), particularly in context with the above field of problems.

Before a specification profile for a communication structure can be realised, it is appropriate to analyse and briefly describe the status of the structural development of healthcare regions, whereby it has to be pointed out that these structure conceptions may only be used as an example, as new IT conceptions have to primarily support highest structure flexibility.

The Tyrolean Concept of an Oncological Network

What particularly shall be highlighted with this example, is that both regional and supraregional remittent structures are, as regards function, directly tied into the diagnostic, therapeutic, and subordinated elements. Therefore, the design of the treatment chain forces the initial contact with the patient as a starting point, even if the initial contact occurs during a public screening event.

On the level of interdisciplinary diagnostics, an integrative conception of diagnostic services (called "medical setup") is established on the basis of diagnostic standards (Todd & Stamper, 1994). Particularly through the additional employment of patient and resource planning instruments, as well as optimised time and process management, brings essential quality and cost advantages. All the more as consistent data structures are accessed within the entire system (Mahr, 1995). As the main result, the interdisciplinary diagnostic responds with further treatment and therapy levels (Berner, Webster, Shugerman, Jackson, Algina, & Baker, 1995), and compiles an optimal (primarily for patients but also for institutions) patient scheduling cooperatively with remitters and clinical facilities.

The intensity of employment of the single facility (stationary or ambulatory) within the entire IT-conception, referring planning, communication, and optimizing tools, too, has to represent the targeted importance within the treatment chain respectively within the research chain.

FUTURE TRENDS

On the whole, it is to be assumed that the question of the interaction and integration of different medical healthcare providers will increase in importance within areas of medical care, as well as within expert chains of treatment. As a consequence, the development of process and information standards has to be developed by the individual specialist associations. Thereby, the scope of intervention potential of single healthcare facilities has to be taken into increasing consideration in order not to restrain the efficiency of a medical facility during a certain step of the treatment chain caused by the organisational and technical paradigm change. This involves integrative and highly standardised IT tools which show a high degree of parameterisation. Beside the impact of more transparent treatment chains, cost efficiency and optimized exploitation of existing health structures, the emerging consistent data will enforce research over the entire treatment chain, as this occurs already (e.g., in nursing with the nursing minimum data set). Minimum data sets for other medical and healthcare providing fields will inevitably arise.

Continuing this approach, the question about the interaction and communication ability of different medical networks with each other arises (i.e., the question about the communication of two or more systems on the meta-level, as well as the operational level respective to integrated patient pathways). In this area, the particular challenge of communication ability is fulfilled, which in turn causes increased research activity in connection with the development of internationally useable terminologies and with the definition of superior standards.

Finally, it is to be assumed that development towards medical networks, aligned processes and real, as well as virtual, medical competence centres will continue to strengthen. In this context, the evolution of organizations' profiles will be of considerable importance, as well as the question of information transfer, knowledge transfer, and quality assurance.

CONCLUSION

All operative healthcare providers in the healthcare branch, be it hospitals, therapy or diagnostic centres, general practitioners or rehabilitation facilities, are subject to considerable cost restrictions while medically expanding both qualitatively and quantitatively.

In the last few years, a higher degree of specialisation of these health facilities has been assessed—considering the entire chain of treatment—which has led to considerable problems in interaction between the health institutions.

Against this background, it appears necessary to create regional cooperations for distinct medical topics. Considering the entire regional treatment chain, the defined care mandates, the areas of competency as well as patient data management, the constitution of a data and communication platform between these institutions on one hand, and shaping this platform dialogue and interaction-capable on the other, challenges organization executives.

In addition, it also seems necessary to integrate optimisation tools, guidelines, and standards into defined interactive and comprehensive processes, which enable methodical and action-orientated resource management supervised by quality standards.

We have detected that patient-orientated communications and operating structures are of strategic importance. The principle question about system functionality and effectiveness of the entire organisation and the organisational structures associated with it is at least of equal importance. In practically all major hospitals, structural organisation occurs according to the historical development of medicine. Even at first glance, it can be recognised that this practice lacks an organisational theoretical basis if the criterion of strategic efficiency should apply (Zulehner, 1999). The reason for this lies in the fact that the abstract levels, according to which the hospitals were traditionally structured, are of a different nature, and are not necessarily compatible with each other. On the one hand, this must lead to double and multiple structures, and on the other, to gaps in the clinical portfolio. In this way, the hospitals are conceptually organised according to an institution or group of institutions, range of services, age of the patients, certain illnesses or diseases, methods and forms of therapy, as well as technical criteria. The medical organisational development of the past decades has of course authority, too, gained from the different speeds at which innovation progresses for the individual lines of expertise. There, it is ensured that single medical sectors that are able to be well combined can develop in the same or similar way, and do not have to experience constraints from other sectors. In the middle and long term, these present organisational structure—if it is not supplemented and reworked in essential areas—will noticeably restrict clinical and economical efficiency. Because of this, it is necessary that new organisational and service structures be defined along side current and continually developing clinical structures for the purpose of a virtual structure or virtual medial centre.

This is particularly characterised by combined standardised processes, a continuous chain of treatment, clear terminology, and uniform quality requirements.

REFERENCES

Bachert, A., & Classen, C.-F. (1995). Wissensbasierte Chemotherapieplanung in der pädiatrischen Onkologie: Ein Beispiel zur Therapieunterstützung. In W. Buchholz, & R. Haux (Eds.), *Informationsverarbeitung in den Universitätsklinika Baden-Württembergs* (pp. 117–124). Heidelberg.

Becker, S.-H., & Arenson, R.-L. (1995). Costs and benefits of picture archiving and communication systems. In J.-H. Van Bemmel, & A.-T. McCray (Eds.), *Yearbook of medical informatics: The Computer-based Patient Record 95* (pp. 345–355). Stuttgart: Schattauer.

Berner, E. S., Webster, G. D., Shugerman, A. A., Jackson, J. R., Algina, J., Baker, A. L., et al. (1995). Performance of four computer-based diagnostic systems. In J.-H. Van Bemmel, & A.-T. McCray (Eds.), *Yearbook of medical informatics: The Computer-based Patient Record 95* (pp. 458–463). Stuttgart: Schattauer.

Bradley, D. A., Williams, G, Brownsell, S. J., & Levi, S. (2002). Community alarms to telecare: The need for a systems strategy for integrated telehealth provision. *Technology and Disability, 14,* 63–74.

Burdea, G. C. (2003). Virtual rehabilitation: Benefits and challenges. *Methods of Information in Medicine, 42,* 519–523.

Cashen, M. S., Dykes, P., Gerber, B. (2004). E-health technology and Internet resources: Barriers for vulnerable populations. *Journal of Cardiovascular Nursing, 19,* 209–214.

Clayton, P.-D. (2001). The state of clinical information systems after four decades of effort. In R. Haux, & C. Kulikowski (Eds.), *Yearbook of medical informatics, digital libraries and medicine 01* (pp. 333–337). Stuttgart: Schattauer.

Coddington, D. (1997). Evaluating integrated healthcare trends: An interview with Dean Coddington. *Healthcare Financial Management, 51*(9), 40–43

Day, K., & Norris, T. (2006). Leadership in times of crises during change due to health IT projects. *Healthcare & Informatics Review Online*. Retrieved February 13, 2008, from http://hcro.enigma.co.nz/website/index.cfm?fuseaction= articledisplay&FeatureID=050906

Francis, S., & Hart, R. (1998). A model for advanced care management and implications for information management. *Healthcare & Informatics Review Online, 12*(2). Retrieved February 13, 2008, from http://hcro.enigma.co.nz/website/ index.cfm?fuseaction=articledisplay&featureid=52

Geißler, B., & Rump, N. (1999). Anforderungskatalog für ein Pflegeinformationssystem unter besonderer Berücksichtigung der klinischen Qualitätsentwicklung. *http://www.PR-Internet.com, Pädagogik, Management und Informatik in der Pflege, 4*, 84–91.

Göbel, G., & Pfeiffer, K.-P. (2001). Qualitätssicherung von patientenorientierten Gesundheitsinformationsnetzen im Internet. In O. Meggeneder, & H. Noack (Eds.), *Gesundheitswissenschaften (16) Vernetzte Gesundheit – Chancen und Risken des Internet für die Gesundheit* (pp. 193–202). Linz: OOEGKK.

Gustafson, D. H., Hawkins, R., Boberg, E., Pingree, S., Serlin, R. E., Graziano, F., et al. (2001). Impact of a patient-centered, computer-based health information/support system. In R. Haux, C. Kulikowski (Eds.), *Yearbook of medical informatics, digital libraries and medicine 01* (pp. 224–232). Stuttgart: Schattauer.

Hain, T. (2002). Improving the quality of health information: The contribution of C-H-i-Q. *Health Expectations, 5,* 270–273.

Hassan, D. (2005). Measuring performance in the healthcare field: A multiple stakeholders' perspective. *Total Quality Management, 16*, 945–953.

Haux, R. (1995). Stand und weitere Entwicklung von Krankenhausinformationssystemen: Die Empfehlungen der Deutschen Forschungsgemeinschaft. In W. Buchholz, & R. Haux (Eds.), *Informationsverarbeitung in den Universitätsklinika Baden-Württembergs* (pp. 20–30). Heidelberg.

Kaloczy, C. (2001). Kommunikation (Krankenhausinformationssystem). In C. Kaloczy, S. Kindermann, & R. Staudinger (Eds.) *Krankenhausorganisation und –finanzierung* (pp. 121–136). Hall in Tirol: Berenkamp.

Kaushal, R., Bates, D. W., Poon, E. G., Ashish, K. J., & Blumenthal, D. (2005). Functional gaps in attaining a national health information network: What will it take to get there in five years? *Health Affaires, 24*, 1281–1289.

Kyriacou, S. K., Davatzikos, C., Zinreich, S. J., & Bryan, R. N. (2001). Nonlinear elastic registration of brain images with tumor pathology using a biomechanical model. In R. Haux, C. Kulikowski (Eds.), *Yearbook of medical informatics, digital libraries and medicine 01* (pp. 429–441). Stuttgart: Schattauer

Leiner, F., Gaus, W., & Haux, R. (1995). *Medizinische dokumentation*. Stuttgart: Schattauer.

Mahr, B. (1995). Principles of open distributed processing in medicine: The Berlin approach. In E. Fleck (Ed.), *Open Systems in Medicine* (pp. 6–23). Amsterdam: IOS Press.

Medical Committee – Royal Netherlands Academy of Arts an Sciences. (1999). *Discipline report on (bio)medical and health sciences research in the Netherlands 1998*. Amsterdam.

Montreuil, B., & Garon, R. (2005). Toward patient centric and distributed healthcare delivery networks. *Clinical and Investigate Medicine, 28*, 351–352.

Ohly, M.-J. (1995). Computer-supported cooperative work. In E. Fleck (Ed.), *Open Systems in Medicine* (pp. 78–87). Amsterdam: IOS Press.

Ölvingson, C., Hallberg, N., Timpka, T., & Lindqvist, K. (2002). Requirement engineering for inter-organizational health information systems with functions for spatial analyses: Modelling a WHO safe community applying use case maps. *Methods of Information in Medicine, 41,* 299–304.

Pernice, A., Doare, H., & Rienhoff, O. (1995). *Healthcare card systems: Eurocards concerted action, results and recommendations*. Amsterdam: IOS Press.

Power, M. (1999). An approach to policy analysis and development of medical informatics. *Methods of Information in Medicine, 38,* 260–264.

Ruggiero, C., Giacomini, M., Sacile, R., Sardanelli, F., Nicolo, G., Lombardo, C. (1995). An environment for computer aided early diagnosis of breast cancer. In International Federation for Medical and Biological Engineering (Eds.), *Proceedings of the Health Telematics `95 – "Health in the Information Society"* (pp. 185–190). Naples.

Sacile, R. (1995). A new approach in home care for oncologic elderly. In International Federation for Medical and Biological Engineering (Eds.), *Proceedings of the Health Telematics `95 – "Health in the Information Society"* (pp. 23–26). Naples.

Schnabel, M. (1996). *Expertensysteme in der medizin*. Stuttgart: Fischer.

Schoop, M., & Wastell, D. G. (2001). Effective multidisciplinary communication in healthcare: Cooperative documentation systems. In R. Haux, & C. Kulikowski (Eds.), *Yearbook of medical informatics, digital libraries and medicine 01* (pp. 379–387). Stuttgart: Schattauer.

Staudinger, R., & Mair, C. (2001). *Rechtstheoretische und rechtstatsächliche Studie zur Frage der Validität krankenanstaltenrechtlicher Bestimmungen am Beispiel der Universitätsklinik Innsbruck*. Innsbruck.

Steffens, J. (1995). Communication infrastructure. In E. Fleck (Ed.), *Open Systems in Medicine* (pp. 104–110). Amsterdam: IOS Press.

Stoop, A. P., & Berg, M. (2003). Integrating quantitative and qualitative methods in patient care information system evaluation: Guidance for the organizational decision maker. *Methods of Information in Medicine, 42,* 458–462.

Tenery, R.-M. (2000). Interactions between physicians and the healthcare technology industry. *Journal of the American Medical Association, 283,* 391–393.

Todd, B.-S., & Stamper, R. (1994). Limits to diagnostic accuracy. In J.-H. Van Bemmel, & A.-T. McCray (Eds.), *Yearbook of medical informatics: Advanced communications in healthcare 94* (pp. 255–270). Stuttgart: Schattauer.

Van den Haak, M., Mludek, V., Wolff, A. C., Bützlebruck, H., Oetzel, D., Zierhut, D., et al. (2002). Networking in shared care: First steps towards a shared electronic patient record for cancer patients. *Methods of Information in Medicine, 41,* 419–425.

Wazana, A. (2000). Physicians and the pharmaceutical industry: Is a gift ever just a gift. *Journal of the American Medical Association, 283,* 373–380.

Yamazaki, S., & Satomura, Y. (2001). Standard method for describing an electronic patient record template: Application of XML to share domain knowledge. In

R. Haux, & C. Kulikowski (Eds.), *Yearbook of medical informatics, digital libraries and medicine 01* (pp. 325–330). Stuttgart: Schattauer.

Zulehner, C. (1999). Digitalisierte pflegedokumentation: Aber wie? http://www.PR-Internet.com*, Pädagogik, Management und Informatik in der Pflege, 3*, 56–66.

KEY TERMS

Directive: Operation and action recommendation with lower commitment than SOPs. Directives describe process paths which should be normally followed during the implementation of the process. Deviations in justified circumstances are allowed.

Effectiveness of Personnel Management: Based on a theoretical maximal quantitative output of an organisation unit in the area of personnel management in connection with costs and efficiency criteria, the effectiveness of personal management is calculated from the difference between the theoretical output and the actual output. It is measured by the different measurement criteria (process quality, data quality, process stability, cost—performance relation, frequency of errors, and so on).

Guideline: Operation and action orientation with lower commitment that directives. Guidelines provide an operation framework in which the process should move. Deviations are not usually sanctioned.

Human Resource Information System (HRIS): Specifically developed software for Human Resource Management (HRM) (e.g., software packets which document, support, and make analysable the digitisable components of HRM). The Human Resource Administration System (HRAS) is an administration-specific subarea of HRIS.

Process Integration: The process integration specifies the degree of the theoretical and actual commitment of processes within structures.

Process Types: Definitions of certain process types. They are distinguishable from another by characteristics in the area of the degree of standardisation and the degree of integration.

Standard Operation Procedures (SOPs): Operation and action instructions. They are there to make sure of the uniform, predictable, and integrated dealing of replaceable process individuals within the framework of working structures.

Virtual Centre: Network of organisational commitments in order to reach cooperation without frontiers between different service units even if they are members of different legal organisations.

IT-Standards and Standardization Approaches in Healthcare

Ali Sunyaev
Technische Universitaet Muenchen, Germany

Jan Marco Leimeister
Technische Universitaet Muenchen, Germany

Andreas Schweiger
Technische Universitaet Muenchen, Germany

Helmut Krcmar
Technische Universitaet Muenchen, Germany

INTRODUCTION

E-health basically comprises health services and information delivered or enhanced through the Internet and related technologies (Eysenbach, 2001). The future healthcare system and its services, enabling e-health, are based on the communication between all information systems of all participants of an integrated treatment. Connecting the elements of each healthcare system (general practitioners, hospitals, health insurance companies, pharmacies, and so on)—even across national boarders—is an important issue for information systems research in healthcare. Current developments, such as upcoming or already-deployed electronic healthcare chip cards (that are to be used across Europe), show the need for Europe-wide standards and norms (Schweiger, Sunyaev, Leimeister, & Krcmar, 2007). In this article, we first outline the advantages of the standards, and then describe their main characteristics. After the introduction of communication standards, we present their comparison with the aim to support the different functions in the healthcare information systems. Subsequently, we describe the documentation standards, and discuss the goals of existing standardization approaches. Implications conclude the article.

BACKGROUND

The advantages of generally accepted standards for the processes in healthcare and the medical market can be summarized as follows (CEN/TC 251 European Standardization of Health Informatics, http://www.centc251.org/; Wirsz, 2000):

- Standards increase competition and reduce costs;
- Standardized products could easily be replaced or updated;
- Standardized products of various suppliers could easily exchange medical information;
- Healthcare institutions are able to iteratively extend their offers/capabilities;
- Standardized products could reduce errors and make healthcare services safer.

Several national and international committees, German as well as European or American (CEN/TC 251 European Standardization of Health Informatics; DIN, http://www.din.de; Integrating the Healthcare Enterprise (IHE), http://www.ihe.net; Integrating the Healthcare Enterprise-Europe (IHE-E), http://www.ihe-eurpoe.org; World Health Organization (WHO), http://www.who.int), have been founded to ensure unified standardization of national and international healthcare systems. Accordingly, there are numerous standardization attempts, which partly correspond to, but also disagree with, each other (Märkle & Lemke, 2002). Two main objectives of these committees can be distinguished: the development of standards for communication, and standards for documentation in healthcare. The former focuses on enabling an efficient and effective combination of medical information systems, in order to enable the exchange of data between different medical information systems (refer to IHE). The latter are supposed to ensure the right interpretation of the content of electronically exchanged information (Haas, 2005).

IT-STANDARDS IN HEALTHCARE

The transmission of data between heterogeneous and isolated medical information systems requires interoperability of systems and data (Hasselbring, 1997). The interoperability on its side is composed of norms, interfaces, and standards—the basis for data exchange and communication between participating applications (refer to World Health Organization (WHO), http://www.who.int). For an overview of common interoperability standards in the healthcare sector, and the graphical classification of its relations, see Figure 1.

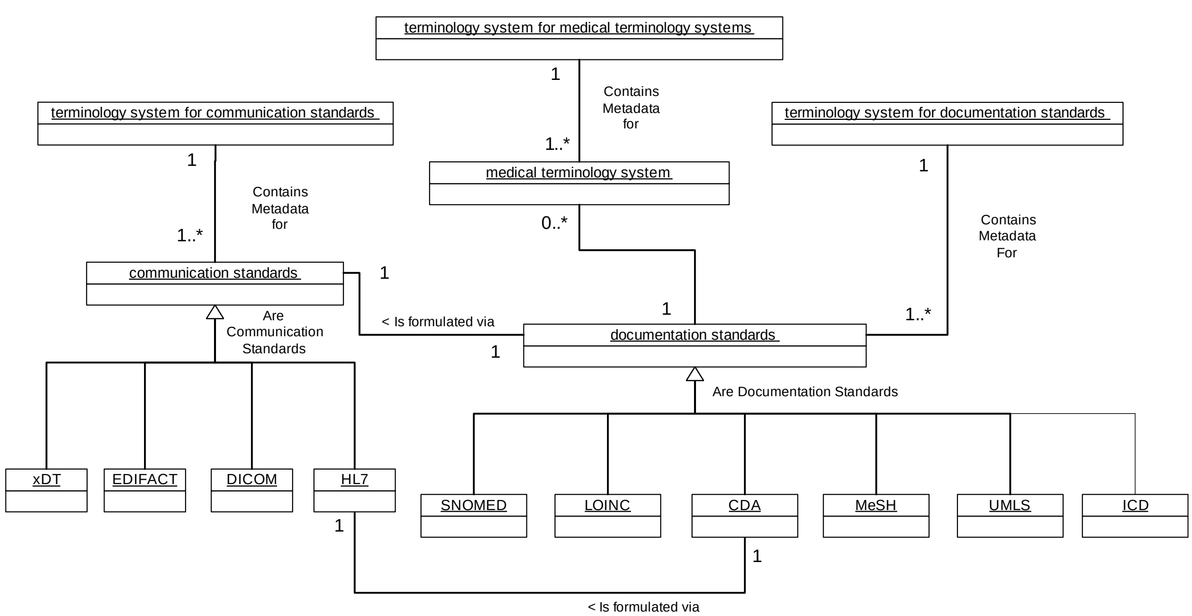
Communication Standards

Communication standards, also called syntactical standards, ensure a correct transmission of medical and administrative data between different information systems. In the clinical area, one can distinguish them worldwide, mainly between the established standards of Health Level 7/Clinical Document Architecture, (HL7/CDA), Digital Imaging and Communications in Medicine (DICOM), and Electronic Data Interchange for Administration, Commerce and Transport (EDIFACT) (Pedersen & Hasselbring, 2004). The reason for their acceptance is mainly the openness of these standards.

Health Level 7 is an international, vendor-independent, and main communication standard in healthcare for the exchange of information between systems and institutions. The Version 2 family of this standard is based on events that trigger the exchange of data. Beginning from Version 3 XML[1], data structure is supported, alleviating the integration of data into information systems, since adequate libraries for XML handling are available. HL7 operates at the application layer of the ISO/OSI reference model (ISO7498-1, International Organization of Standardization (ISO), 2006). The Clinical Document Architecture (Health Level Seven Inc., http://www.hl7.org/) amplifies the HL7-standard with the description of the structure and the contents of clinical documents (e.g., discharge summaries and progress notes), based on an XML-format. HL7/CDA also offers a model for the exchange or the common use of information, and the option to individually re-use this information (Dolin, Alschuler, Boyer, Beebe, Behlen, & Biron, 2006).

Digital Imaging and Communications in Medicine (National Electrical Manufacturers Association (NEMA) , http://medical.nema.org) is an open standard

Figure 1. Medical communication and documentation standards



for the exchange of images in healthcare. Besides the image, it can also contain additional meta-information, such as the patient's name, date of admission, device parameters, or attending physician. The standard lists the data fields (e.g., images, diagnosis, patients, studies, series, and so on), functions of network services, as well as syntax and semantics for the commands and messages. DICOM can store the images with or without loss of information, in accordance to TIFF and JPEG formats, and enables electronic archiving of images in all medical information systems.

Electronic Data Interchange For Administration, Commerce and Transport (EDIFACT, 2006) standardizes the formats for the electronic exchange of administrative data amongst others in healthcare information systems. Orders, calculations, and payment orders in the health service are often based on it. EDIFACT is, unlike the other introduced communication standards, not specifically designed for the healthcare field—it is used globally by business partners of different kinds of business areas to exchange accounting-relevant data.

In the sector of general practitioners in Germany, there are additionally specific standards, so called xDT-standards (KBV, 2006). xDT is a set of several German standards, which are supposed to simplify the communication and the data exchange between healthcare providers (i.e., physician's offices, hospitals, and so on) and health insurance companies. As a result, xDT aims at the reduction of waste of paper by applying information technology using the following data formats:

- **ADT:** AbrechnungsDatenTransfer (invoice data for medical services)
- **AODT:** AmbulantesOperierenDatenTransfer (securing the quality of out-patient surgery)
- **BDT:** BehandlungsDatenTransfer (for the communication between general practitioners)
- **KVDT:** KassenaerztlicheVereinigungDatenTransfer (standard for communication between physician's offices and health insurances)
- **LDT:** LaborDatenTransfer (communication with laboratories)
- **GDT:** GeraeteDatenTransfer (communication with medical devices)
- **ODT:** OnkologischerDatenTransfer (oncological data carrier for the documentation of tumor information)

Even though all of these xDT-standards serve different purposes, they are structured according to the same basic rules and principles.

There are currently efforts to harmonize the introduced communication standards on an XML-basis: HL7/CDA incorporates already the XML-format for structuring clinical documents; additionally, there are also XML-based specifications for the DICOM standard (refer to NEMA), for the xDT-standards (Bundesministerium für Gesundheit und Soziale Sicherung, 2004), and for EDIFACT (EDIFACT, 2006). The advantages of introducing XML-based data structures can be described as follows:

- Automatic processing can be simplified when using established XML programming libraries.
- XML is an open and platform-independent standard.
- XML can be read and understood by users.

The introduced communication standards support different functions in healthcare information systems (Table 1: "x" connotes that the standard supports the function, "-" connotes the opposite case). Such a use of miscellaneous standards in several healthcare service sectors is the most current problem in electronic communication in medicine (Lenz, Beyer, Meiler, Jablonski, & Kuhn, 2005).

Documentation Standards and Standardization Approaches

For the right interpretation of the content of the electronically exchanged messages between different medical information systems, there are terminology and documentation standards, called semantic standards (see also Figure 1). The semantic standards are responsible for the conversion of encoded medical data, in order to, for example, correctly identify the diagnosis. Logical Observation Identifiers Names and Codes (LOINC®, http://www.regenstrief.org/loinc) offers a multidimensional terminology system for clinical laboratories. It allows detailed descriptions of medical circumstances for almost each clinical problem to be solved automatically.

Classifications are medical systems of concepts that are used for medical documentation. In comparison to terminology systems, classifications are less complicated, and thus mostly one-dimensional. "International

Table 1. Communication standards and their functions

	Medical communication standards			
Functions in health care	DICOM	xDT	EDIFACT	HL7/CDA
HIS	x	x	x	x
RIS	x	-	-	x
PACS	x	-	-	x
MPI (Master Patient Index)	-	x	-	x
graphical diagnosis	x	-	-	x
archiving	x	x	-	x
diagnosis comments	-	x	-	x
image documentation	x	-	-	x
report comments	-	-	-	x
video documentation	-	-	-	x
patient registration	-	x	-	x
ERP (Electronic Health Records)	x	x	x	x
invoicing	-	x	x	x
prescriptions	-	-	-	x
data transformation	-	x	-	x
emergency data	-	-	-	x
physician's practices	-	x	x	x

HIS - Hospital Information System,
PACS - Picture Archiving and Communication System,
RIS - Radiology Information System.

Classification of Diseases" (ICD) (DIMDI, http://www.dimdi.de; refer also to WHO) is an example for such a classification of illnesses and related health issues.

Systemized Nomenclature of Medicine (SNOMED, http://www.snomed.org/index.html) is a nomenclature. Nomenclatures are combinations of terms that are based on determined concept orders. SNOMED CT covers the arrangement of a unified terminology for expressions in the medical field, and supports the languages English, German and Spanish.

The thesaurus Medical Subject Headings (MeSH) (United States National Library of Medicine, 2006) is an extension of a nomenclature. MeSH mainly enables indexing international publications—such as journals, articles, and books—for the United States National Library of Medicine (NLM). Enriching a thesaurus with semantic and linguistic information yields a metathesaurus. The "Unified Medical Language System" (UMLS, http://www.nlm.nih.gov/research/umls) is a metathesaurus that tries to integrate all important medical terms in only one term, and to represent all possible term relations accordingly.

Because of the existence of different documentation standards, a lot of incompatibilities occur, such as if different users use different terminology systems for the documentation (McDonald, 1997). The complete interoperability, the consent identification concerning the meaning of the medical data and meaningful

Table 2. Overview of current standardization approaches in healthcare

Standardization approaches in healthcare	Characterisation and Goal	Approach and assignment of internationally accepted standards
Integrating the healthcare enterprise (IHE) (Hornung, Goetz, & Goldschmidt, 2005)	IHE is an international initiative, driven by health-care professionals and industry, to improve the communication of medical information systems and the exchange of data.	IHE's approach for the information integration is based on the propagation and integration/usage of HL7 and DICOM standards. IHE promotes and advances these standards as a suggestion for standardizing bodies.
Professionals and citizens network for integrated care (PICNIC) (Danish Center for Health Telematics, 2003)	PICNIC is a European project of regional healthcare providers in a public-private partnership with industry to develop new healthcare networks, and to defragment the European market for healthcare telematics.	The development is an open source model, an open and interoperable architecture with exchangeable components (aim is an easy and simple integration of external products). All components must be based on established standards, such as HL7/CDA.
Distributed healthcare Environment (DHE, http://www.gesi.it/dhe)	DHE has been created as a part of previous development projects (e.g., HANSA-Project (http://www.ehto.org/ht_projects/initial_project_description/hansa.html) for the integration of clinical information systems) to build a platform to both integrate legacy healthcare systems, and support the development of new ones.	DHE is an open middleware, which is based on the "Healthcare Information System Architecture" (HISE, CEN/TC 251 European Standardization of Health Informatics) (Scherrer & Spahni, 1999). DHE tries to specify generic healthcare services, which are based on predefined database-schemata, and offers them on a common platform (e.g., with HL7-interfaces).
Open electronic health record (The openEHR foundation, http://www.openehr.org)	The openEHR foundation is dedicated to develop an open specification and implementation for the electronic health record (EHR). openEHR advances the experiences of Good Eurpoean Health Record-Projects (Blobel, 2006; GEHR, http://www.chime.ucl.ac.uk/work-areas/ehrs/GEHR/) in England and Australia.	The project works closely with standards (e.g., HL7). However, it does not adopt them verbatim, but tests, implements, and improves their integration and application, while giving feedback to the standardizing bodies.
Standardization of communication between information systems in physician's offices and hospitals using XML (Gerdsen, Müller, Bader, Poljak, Jablonski, & Prokosch, (2005); SCIPHOX, 2006)	SCIPHOX is a German initiative with the aim to define a new common communication standard for ambulant and inpatient healthcare facilities.	The basis for the information exchange is the XML-based HL7/CDA standard. SCIPHOX adapts and improves this global standard for local (German) needs.
www.akteonline.de (Schwarze, Tessmann, Sassenberg, Müller, Prokosch, & Ückert, 2005; Ückert, Görz, Ataian, & Prokosch, 2002)	akteonline.de is German state-funded project to develop a Web-based electronic healthcare record.	akteonline.de developed dynamic Web pages, which can be accessed via Internet, and look similar to physicians and hospital software. The project is based on the common communication standards (DICOM and HL7/CDA).
All introduced initiatives, with the exception of SCIPHOX, defined and developed an "Electronic Healthcare Record" (EHCR) (Tang, Ash, Bates, Overhage, & Sands, 2006)		

cooperation of functions, are the primary ones of the main challenges for the near future of development of healthcare information systems (Frist, 2005). There are many different approaches to standardize the electronic communication and documentation in healthcare to achieve the interoperability. Some of the promising standardization initiatives are summarized in Table 2.

The analysis described in Table 2 shows that the existing standardisation approaches for interoperability in healthcare can be divided in two opposing groups (Märkle & Lemke, 2002). On the one hand, there is basically industry that propagates and adopts proprietary standards like HL7 and DICOM. On the other hand, there are generally university- and state-funded initiatives which use and support open standards as initiated by European Committee of Standardization (CEN/TC 251 European Standardization of Health Informatics). This difference is also based regionally—the former are mainly located in the U.S., whereas the latter initiatives come from Europe. For the future healthcare system, it remains to be aspired that both camps learn from each other and may grow together, and with it, the vision of complete interoperability in the healthcare sector can become true. Nevertheless, they all provided important issues and contributed indispensable knowledge in attaining the present status.

IMPLICATIONS

Standards in healthcare are supposed to enable an efficient and effective communication of distributed and isolated medical information systems. Unfortunately, the current situation is not satisfying in terms of interoperability. Until there is no generally accepted standardization for syntactic as well as semantic standards in the healthcare sector, all advantages of standardized solutions listed in the introduction cannot be exploited (Haux, 2006). To improve health services and reduce costs, the problem of interoperability (standardized software solutions) has to be solved. Such interoperable information systems can influence the quality of healthcare services and the adjunctive costs very positively (Berger & Partner GmbH, 1997; Warda & Noelle, 2002).

The issue of interoperability as major driver for efficiency in healthcare requires efforts as to harmonization of standards, especially on the semantic level. The existing and introduced communication standards are able to deal with and connect different medical information systems. But a great percentage of the potential benefits still remain useless without prior harmonisation of standards on a semantic level. Future efforts from research and practice should focus on this topic.

REFERENCES

Berger, R. & Partner GmbH – International Management Consultants. (1997). *Telematik im gesundheitswesen: Perspektiven der telemedizin in Deutschland*. Im Auftrag des Bundesministeriums für Bildung, Wissenschaft, Forschung und Technologie und in Zusammenarbeit mit dem Bundesministerium für Gesundheit. Retrieved March 26, 2008, from http://www.wernerschell.de/Medizin-Infos/Telemedizin/Telematik%20im%20Gesundheitswesen9999.htm

Blobel, B. (2006). Advanced and secure architectural HER approaches. *International Journal of Medical Informatics, 75*, 185–190.

Bundesministerium für Gesundheit und Soziale Sicherung. (2004). Standards und Initiativen im Gesundheitswesen - eine evaluierende Übersicht und Empfehlung. Retrieved March 26, 2008, from http://www.dimdi.de/dynamic/de/ehealth/karte/downloadcenter/technik/rahmenarchitektur/telematik_rahmenaktuell/b4h_standards_v1-1.pdf

Danish Center for Health Telematics. (2003). *PICNIC (professionals and citizens network for integrated care) Web site*. Retrieved February 2, 2008, from http://www.medcom.dk/picnic

Dolin, R. H., Alschuler, L., Boyer, S., Beebe, C., Behlen, F. M., Biron, P. V., et al. (2006). HL7 Clinical Document Architecture, Release 2. *Journal of the American Medical Informatics Association, 13*, 30–39.

EDIFACT. (2006). *United Nations economic commission for Europe*. Retrieved February 2, 2008, from http://www.unece.org/

Eysenbach, G. (2001). What is e-health? *Journal of Medical Internet Research, 3*(2), e20.

Frist, W. H. (2005). Healthcare in the 21st century. *The New England Journal of Medicine, 352*, 267–272.

Gerdsen, F. M., Sascha, Bader, E., Poljak, M., Jablonski, Stefan, Prokosch, Hans-Ulrich. (2005). *Einsatz von CDA/SCIPHOX zur standardisierten Kommunikation medizinischer Befunddaten zwischen einem Schlaganfall-/Glaukom-Screening-Programm und einer elektronischen Gesundheitsakte (EGA)* . Paper presented at the Proceedings of Telemed 2005, Berlin.

Haas, P. (2005). *Medizinische informationssysteme und elektronische krankenakten*. Berlin, Heidelberg, New York: Springer.

Hasselbring, W. (1997). Federated integration of replicated information within hospitals. *International Journal on Digital Libraries, 1*(3), 192–208.

Haux, R. (2006). Health information systems—Past, present, future. In *International Journal of Medical Informatics*, *75*(3-4), 268-281.

Hornung, G., Goetz, C. F. -J., & Goldschmidt, A. J. W. (2005). Die künftige telematik-rahmenarchitektur im gesundheitswesen: Recht, technologie, infrastruktur und ökonomie. *Wirtschaftsinformatik, 46*(3), 171–179.

KBV. (2005). Kassenärztliche Bundesvereinigung Deutschland. xDT - Synonym für elektronischen Datenaustausch in der Arztpraxis. Retrieved February 2, 2008, from http://www.kbv.de/ita/4274.html

Lenz, R., Beyer, M., Meiler, C., Jablonski, S., & Kuhn, K. A. (2005). Informationsintegration in gesundheitsversorgungsnetzen: Herausforderungen an die informatik. *Informatik-Spektrum, 28*(2), 105–119.

Märkle, S., & Lemke, H. U. (2002). Die elektronische patientenakte—ist eine standardisierung in sicht. In *Tagungsband der Jahrestagung des VDE,* Band 2, 69-74.

McDonald, C. J. (1997). The barriers to electronic medical record systems and how to overcome them. *Journal of the American Medical Informatics Association, 4*(3), 213–221.

Pedersen, S., & Hasselbring, W. (2004). Interoperabilität für informationssysteme im gesundheitswesen auf basis medizinischer standards. *Informatik - Forschung und Entwicklung, 18*(3–4), 174–188.

Scherrer, J. -R., & Spahni, S. (1999). Healthcare information system architecture (HISA) and its middleware models. In *Proceedings of AMIA Symposium* (pp. 935–939).

Schwarze, J.-C., Tessmann, S., Sassenberg, C., Müller, M., Prokosch, H. -U., & Ückert, F. (2005). Eine modulare gesundheitsakte als antwort auf kommunikationsprobleme im gesundheitswesen. *Wirtschaftsinformatik, 47*(3), 187–195.

Schweiger, A., Sunyaev, A., Leimeister, J.M., & Krcmar, H. (2007). Information systems and healthcare XX: Toward seamless healthcare with software agents. In *Communications of the Association for Information Systems, 19*(33), 692-709.

Standardized Communication of Information Systems in Physician Offices and Hospitals using XML (SCIPHOX). (2006). Arbeitsgemeinschaft SCIPHOX GbR mbH c/o DIMDI. Retrieved March 26, 2008, from http://www.sciphox.de

Ückert, F., Görz, M., Ataian, M., & Prokosch, H. -U. (2002, August 25–29). *Akteonline - an electronic healthcare record as a medium for information and communication.* Paper presented at the Medical Informatics Europe, Budapest.

United States National Library of Medicine, N. I. O. H. (2006). MesH - Medical Subject Headings. Retrieved February 2, 2008, from http://www.nlm.nih.gov/mesh/

Warda, F., & Noelle, G. (2002). *Telemedizin und e-health in Deutschland: Materialien und empfehlungen für eine nationale telematikplattform (1 Ed.)*. Videel OHG.

Wirsz, N. (2000). IT-standards im gesundheitswesen. *Electromedica, 68 Heft 1*.

KEY TERMS

Classification: Given representation of abstract concepts, which are assigned to a set of categories.

Communication (Syntactic) Standards: Communication standards ensure a correct transmission of electronic data between different information systems, in terms of a particular syntactical format.

Documentation (Semantic) Standards: Documentation standards ensure the right interpretation of encoded medical data.

Interoperability: Denotes the ability of different, independent, and heterogeneous information systems to exchange electronic data for their further processing.

Metathesaurus: A thesaurus enriched with additional semantic and linguistic information.

Nomenclature: An accumulation of principles for the consistent naming of objects.

Thesaurus: A given vocabulary with related terms.

ENDNOTE

1 The standard extensible mark-up language XML, as defined by the World Wide Web Consortium (W3C), is used for the creation of machine- and human-readable documents in the shape of a tree structure, and is hereby defining the structure of these files. An actual document requires the definition of several details, particularly the definition of elements of structure and their arrangement within the document tree.
XML is a subset of the Standard Generalized Mark-up Language (SGML). It is also a metalanguage, which is capable of defining numerous amounts of different mark-up languages, which are still closely related in their basic structure. In order to describe the structure of XML documents, a so-called "scheme language" is used. The two most important scheme languages are the Document Type Definition (DTD) and the XML schema.

J

Judging the Value of Mobile Healthcare Solutions for Improving Outpatient Adherence

Mihail Cocosila
McMaster University, Canada

Norm Archer
McMaster University, Canada

Yufei Yuan
McMaster University, Canada

INTRODUCTION

A serious problem in today's health care is the growing incidence of chronic diseases and conditions with dire human and financial consequences. For instance:

- The worldwide burden of chronic conditions will increase from approximately 600,000 *Disability Adjusted Life Years (DALYs),* that is, lost years of healthy life caused by premature death or disability, to over 900,000 *DALYs* worldwide during the period 1990-2030 (WHO, 2003);
- The care of the 90 million Americans with chronic conditions consumes more than $1,000 billion each year, 75% of national medical costs (Chin, 2002); and,
- The high incidence of chronic illnesses is one of the causes for increasing health care costs in Canada, with total health care costs reaching about 10% of the gross domestic product (GDP) (Coambs, 2002).

The implications of chronic diseases and conditions, associated with a growing elderly population that needs more medical attention (Siau, Southard, & Hong, 2002), represent a supplementary challenge for health care systems. These are already facing an increasing divergence between population demands for better care and the resources available to meet these demands (Cowling, Newman, & Leigh, 1999).

Managing some chronic illnesses in outpatient conditions may provide solutions to alleviate some of the pressure on the health care system. Keeping patients out of hospitals while providing them the best treatment conditions possible could be a win-win situation for both people and society. However, there is a serious obstacle to this solution: patient adherence to treatment is no more than 50% on average, with dire human and financial consequences (Dezii, 2000; WHO, 2003).

The problem of low adherence in outpatient conditions on one side and the ubiquity of mobile communications and data services on the other side suggests an innovative idea: improve some forms of outpatient adherence through the use of mobile information technology (IT). This idea springs from the general approaches recommended for improving adherence in relevant medical literature (Haynes, Yao, Degani, Kripalani, Garg, & McDonald, 2005; McDonald, Garg, & Haynes, 2002). Although the philosophy of using the newest technology for promoting wellbeing may seem attractive, various human, technological, and financial issues must be considered first (Cocosila & Archer, 2005b).

On the adoption of mobile technology, an important aspect that has not been addressed adequately in the literature is the business perspective. The objective of this study is to introduce a framework for determining the business case for mobile information technology (IT) solutions that will help to increase outpatient adherence.

BACKGROUND

Chronic diseases and conditions are either infectious (e.g., HIV-AIDS, tuberculosis) or non communicable (e.g., diabetes, hypertension, or asthma). They have several distinct features:

- Duration (permanent or necessitating a long period of care);
- Consequences (caused by nonreversible pathological deterioration and associated with remanent disability); and,
- Treatment (necessitating multidisciplinary management and special conditions and training for patient rehabilitation) (WHO, 2003).

Because of the long-term treatment required, maintaining high adherence becomes a fundamental issue, especially in outpatient conditions where patient contact with health care providers is infrequent. Adherence is a complex phenomenon, still little understood, being simultaneously influenced by several categories of factors (e.g., sociodemographic, economic, medical condition-related, therapy-related, health care team, and system-related, and, above all, patient-related) (WHO, 2003). Since "current methods of improving adherence for chronic health problems are complex, labor-intensive, and not predictably effective" as McDonald et al. (2002, p. 2868) pointed out, a newer approach that uses mobile IT solutions may be justifiable. It must be recognized that the immediate influence of adherence-improving interventions may be small. However, the long-term effects on patients, their social environment, and society in general could mean a significant gain.

Mobility in the context of communication and information "should not be understood simply as a new distribution channel, a mobile Internet, or a substitute for PCs" (Yuan & Zhang, 2003). A mobile IT solution is justified only by the particular *location* of the user, who needs an information or communication service, or by the *utility* or *urgency* of that service (Mennecke & Strader, 2003). All three elements justify the use of mobile solutions to improve outpatient adherence in self-management programs for chronic illnesses. A further important characteristic of any mobile solution is the *identification of a unique person,* as stressed by Junglas and Watson (2003). Thus any communication between a mobile device and health providers would be automatically linked with the mobile device owner's identity.

It is unrealistic to expect mobile solutions to improve adherence automatically. Mobile solutions should aim at facilitating the communication and exchange of data between the patient and the health care provider. Consequently, to maximize the chances of success, any mobile solution aimed at enhancing outpatient adherence should target several adherence factors within the complexity of clinical interventions, where improving the patient-health provider relationship is a key factor. Starting from the recommendations of pertinent medical research (Haynes, McDonald, & Garg, 2002), six possible mobile health care solutions (monitoring of health parameters, reminding about taking a medication or performing a behavioural change, consulting with health providers, receiving support from family and peers, keeping informed, and being educated about the disease and treatment) and their envisioned benefits have been described in previous work (Cocosila & Archer, 2005a). These solutions address adherence factors and extend existing ways of communication and data exchange between outpatients and the health care system (e.g., mail, landline telephone, or the Web), encompassed under the definition of telemedicine (Bashshur, Shannon, & Sapci, 2005).

Existing Applications

Advances in wireless and Internet technology have already made possible the development of new health care services. Two of the most popular are remote monitoring and disease management for patients in homecare. A synthesis of some existing or proposed mobile applications, as mentioned in the literature, is presented in Table 1.

Trials of distance monitoring of blood glucose or blood pressure with the help of a computer showed improved outcomes and efficiency together with more active participation of patients (Balas & Iakovidis, 1999). Extrapolating from the success of preliminary results with the *Telephone-Linked Care* technology, Friedman, Stollerman, Mahoney, and Rozenblyum (1997, p. 424) suggest that "the market demand for technology-based delivery systems used by patients in their homes will be strong." Husemann (2004) sees a future in the emerging field of information-based medicine for solutions like monitoring vital measurements on-the-go or for pill dispensers reminding the patients to take the appropriate medicine.

Electronic monitoring is constantly expanding and innovative devices and applications are becoming increasingly available for a variety of diseases. In particular, analysts foresee wireless medical devices helping chronically ill outpatients with asthma or diabetes (Duan, 2003; Hoise, 1999) and distance monitoring of diabetic patients has already proven to ameliorate the

Table 1. Existing or proposed applications of mobile health care addressing outpatient adherence

Study	Type of intervention	Stage of mobile IT implementation	Disease
Husemann (2004)	Monitoring of vital measurements	Proposed	Not specified
Duan (2003)	Not specified	Proposed	Asthma, diabetes
Hoise (1999)	Not specified	Proposed	Asthma, diabetes
Weingarten et al. (2002)	Monitoring	Not specified	Diabetes
Wexler (2001)	Monitoring health parameters, alerting	Proposed	Diabetes
Dunbar et al. (2003)	Reminding about taking medication and behavioural changes	Implemented	HIV
Conlin (2000)	Patient information and communication with physicians	Implemented	Not specified
Wirelessnewsfactor (2000)	Reminding about taking medication	Implemented	Not specified
Huseman (2004)	Monitoring the taking of medication	Implemented	Not specified
Maglaveras et al. (2002)	Monitoring	Not specified	Not specified
ProWelness (2004)	Monitoring	Implemented	Diabetes
Van Impe (2001)	Monitoring	Implemented	Not specified
Bludau (2003)	Monitoring blood pressure, pulse rate, or temperature in hospitals	Implemented	Not specified
Poropudas (2001)	Monitoring	Not specified	Not specified

clinical outcomes (Weingarten et al., 2002). Wexler (2001) reports on a development by the *IMetrikus Company* that allows diabetics to test their health parameters and upload the results to a Web site. The future could, reportedly, bring a Web-free solution using a direct connection between the glucose monitor and a cell phone, without patient intervention. Furthermore, a health care alert triggering feature could also be incorporated. A study of 25 HIV-positive participants included sending them between three and eight text message reminders per day about taking their medications. Patients reacted positively with resulting behavioural changes. They "expressed high satisfaction with the messaging system and reported that it helped with medication adherence" (Dunbar et al., 2003, p. 11).

Medtronic CareLink™Network has been promoting a remote Web-based monitoring service for patients with implanted cardiac devices. Patients can send timely information to their clinics from regular phones (Medtronic, 2004). *WellMed Inc.* developed a wireless application allowing patient data transmission (e.g., health or insurance information) and communication with health care providers (e.g., physicians) (Conlin, 2000). *SmartMeds.com* and *AT&T Wireless* from the USA. have jointly developed a wireless solution for reminding patients to take prescribed medications and allowing patient feedback about complying with the reminders (Conlin, 2000; Wirelessnewsfactor, 2000). *IBM* researchers pioneered an information technology solution to track vital health signs and send them to remote sites. The device includes a pillbox and a wristwatch-like blood pressure monitor able to communicate via a Bluetooth short-range radio connection to a cell phone. The phone, in turn, sends the data to health care personnel. Only authorized personnel are able to view the patient data. This technology is thought to reduce the number of check-ups, hospitalization periods, and patient health risks (Husemann, 2004).

In recent years, significant work has been underway in Europe on Internet or wireless remote monitoring. Maglaveras et al. (2002) advocate the "necessity for restructuring medical knowledge for education delivery to the patient" and exemplifies this by the novel idea of a contact centre enabled by wireless technology to manage the home care of patients remotely. Patients using *ProWelness,* a Web-based information system for remotely managing several chronic diseases including diabetes, are able to record specific disease management activities daily by fixed phones or cell phones. The philosophy behind the system is the two key elements essential for the success of a chronic disease (e.g., diabetes) management: "careful self-care and good communication with the care team" (ProWellness, 2004). Following the implementation of a 3 year project financed by the *British Engineering and Physical Sciences Research Council*, doctors in the UK were able to monitor remotely the health condition of patients by analyzing medical data gathered by a device communicating through a cell phone (Van Impe, 2001). Bludau (2003) describes a platform for patient home monitoring by hospital physicians. The solution uses Bluetooth-enabled modular sensors and cell phone technology to send data such as blood pressure, pulse rate, and temperature. *Arbonaut*, from Finland, and *Virtual Medical World Solutions*, from the UK, have been developing a platform for real-time remote monitoring. This solution targets patients with a "stable medical condition that allows a near normal life but may suddenly deteriorate and put life at risk" (Poropudas, 2001, p. 1).

MOBILE SOLUTIONS TO IMPROVE OUTPATIENT ADHERENCE: A BUSINESS EVALUATION

A careful evaluation of the business value proposition for mobile applications in the health care sector is necessary in order to assess the risks associated with deployment. Business evaluation is a step-by-step process, as described by Archer (2004) and depicted in Figure 1. In Figure 1, solid arrows represent information and process flows, whereas the dashed arrow represents moderating influences. The business process is described in more detail in the following.

Business Goal or Operational Imperative

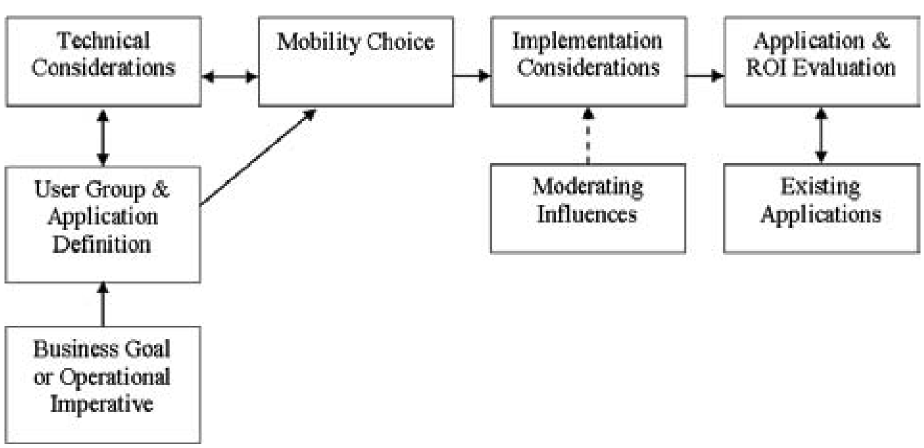
The pillars supporting the option for a mobile health care solution must be emphasized from the very beginning. They should not be a mere justification for the deployment of a particular application or technology. The main reasons for using mobile solutions to improve outpatient adherence should be present for the three most important stakeholders:

- Better quality of care and life for patients (e.g., less medical, personal, and social consequences of their diseases) (WHO, 2003);
- Better use of health providers' professional time and work resources (e.g., less waste for unnecessary diagnostic tests, appointments, and follow-ups) (Weiner, 2004); and,
- Increased savings of human and financial efforts for the health care system and society (e.g., better redistribution of resources to patients needing them most) (Dezii, 2000).

User Group and Application Definition

Mobile solutions addressing nonadherence in outpatient conditions have as users a group of patients with an active lifestyle (i.e., going to work, school, or various locations) treated through long-term care in out-of-hos-

Figure 1. Process framework for mobile solutions to improve outpatient adherence—business case evaluation—adapted from Archer (2004)



pital conditions (i.e., primary and/or home care). The most suitable patients for these types of applications are those suffering from chronic diseases or conditions that require frequent measuring and evaluation of health parameters as well as communication (e.g., seeking advice from the responsible nurse) and data exchange (e.g., sending glucose test results to a monitoring centre for recording and comparison purposes) with health providers.

Technical Considerations

Deploying a new technology application poses several types of technical issues of which the most important are: integration with existing infrastructure, consideration of application and device capabilities, and harmonization with systems already in place (Turisco & Case, 2001). Mobile solutions have in addition several types of technical inhibitors, compared to the wired Internet, due mainly to limitations of the application and mobile device: limited connectivity, low bandwidth, limited input/output capabilities, small memory and low processor capability, and inability to support face-to-face interactive dialogue between patient and provider for observation and examination.

Security and privacy are an additional concern for mobile applications involving patients or patient data. Although there is still a widespread perception that wireless communications are less secure then their wired equivalents, appropriate encryption measures are available and in use for health care information transmission and storage regulated by the Health Insurance Portability and Accountability Act (HIPAA) in the USA and the Personal Information Protection and Electronic Document Act (PIPEDA) in Canada, for instance.

Mobility Choice

An initial question is to ask whether the solution for outpatient adherence should be 100% wireless. Mobility does not automatically imply wireless access. Wireless is only one technology (e.g., cell phones) that enables mobility by the use of specific wide area networks. *Wireless technology is best suited for enabling the required features in most of the previously identified interventions to improve adherence*. Nurturing patient self-confidence and perceived protection, realizing a virtual permanent connection with health providers as well as with social support groups, receiving timely information of interest, alerting the health care system when some indicators exceed the normal limits, and asking for assistance and fast intervention in emergency situations cannot be effective without a wireless connection between the patient and the health care system. However, this linkage must be selective and gradual, based on segmentation and prioritization of patient conditions, so as to not overload the health care system with unnecessary information when patient situations do not require intervention. Thus, "the system would have improved efficiency in managing the patients' uncommon conditions" (Cocosila, Coursaris, & Yuan, 2004).

For other self-management features, it is arguable whether a wireless connection is necessary. With the goal of improving patient quality of care, while offering savings of health care resources, some other features such as constantly monitoring or reminding patients to take the prescribed medication, for adhering to a prescribed diet, or performing self-tests at the required time may be fulfilled by a portable but not wireless application implanted in the patients' mobile devices. Wireless networks and always-on connectivity should only be used for remotely contacting health care personnel in out-of-normal situations, as well as for alerts and emergency functions. For normal situations, the application might only use the built-in capabilities of the mobile device. However, this would prevent real-time access to the health care system, providing information on patient health and medication state, and it would be unable to provide useful provider feedback, which is known to encourage patient adherence. In order to assess the balance between functionality and cost effectiveness, a more complete analysis of business aspects is necessary to find the best compromise for patient care.

Implementation Considerations

Implementing mobile solutions to improve outpatient adherence would have to take into account several issues (Tarasewich, 2003; Venkatesh, Ramesh, & Massey, 2003): system design, device and vendor choice, deployment planning, installation, support and training, prototype and usability testing, and installing of control mechanisms for the deployment stage and for future use. For maximizing the chances for success of adherence improving initiatives, it is advisable to use existing wireless networks but to allow flexibility in the choice of device. Mobile devices should best fit the previously identified groups of user needs. For instance, for active people already using or familiar with cell phones, it would be rational to use this type of common mobile device (with text messaging capabilities) for health care activities since cell phones would also be used frequently by patients for other activities. Regular cell phone use has social importance, since it diminishes the risk of indicating a chronic health problem to colleagues and friends. A cell phone may also increase confidence in using the latest technology and giving better social status. On the other hand, for other people such as the elderly, or for those living mostly in a fixed location, some more specific and easier to use devices could be considered. In addition, for these people the possibility of integration of mobile devices with self-test devices (e.g., blood pressure monitors) for automatically measuring and transmitting data to health care providers would likely be relevant.

Moderating Influences

Moderating influences refer to factors that affect successful implementation and are similar to any application that is new to an organization. Some critical success factors refer to: top management support, a well-planned implementation process, setting up an appropriate mobilization team, a significant degree of e-business implementation, an appropriate organizational culture, choosing the right Value Added Retailer (VAR), establishing communication channels, providing education and training, ensuring comprehensiveness of system functions, and setting up appropriate control mechanisms (Luarn, Lin, & Lo, 2003). In addition to these is the full understanding and matching of user needs. This question is fundamental for health care applications since complex issues such as the social image of mobile application users could have a decisive influence. For instance, using an advanced mobile device and application could improve social image, thus contributing to the acceptance of the application. However, a contrary view is that a mobile application reminding a user to take a certain medication or to perform self-tests may be an embarrassing indication of a chronic illness to colleagues and friends.

Mobile Application and Return on Investment Evaluation

An essential condition prior to evaluation of the Return On Investment (ROI) is that the business goals are properly addressed by the project (Archer, 2004). Financial measures upon which ROI is based include the cost of software, hardware, implementation, ongoing maintenance, and operating costs. Savings include quantifiable (or tangible) and nonquantifiable (or nontangible) elements. Quantifiable elements refer to medical savings, and other (personal and social) savings. Nonquantifiable benefits may be at or above the importance of quantifiable benefits: improving outpatient quality of care and of life (including better integration and social image), prolonging lives, or more professionally rewarding work for health care providers.

Two basic approaches are possible in implementing mobile solutions for outpatient adherence: a *minor change* (that would be an improvement in terms of quality and efficiency of the existing system where the main point of contact for outpatients is represented by homecare nurses), or a *major change* (such as an important advance in existing systems, such as the involvement of physicians in remote consultations). Although improving productivity in health care may translate to reductions in medical staff because of their substitution by technology, this work does not promote such a perspective because it may generate unease among health care workers (Jacobs, 1999). Rather, it endorses the view of Montori, Helgemoe, Guyatt, and Dean (2004) where, even in the most conservative situation when adherence is not improved and there are not savings in health care worker time, mobile solutions for outpatients may help in redistributing health care provider time and work resources more fairly to the patients needing them most. A final step in the evaluation of ROI is to compare the mobile solution with the existing process and its associated business model so as to make sure that radical change in the existing system is better than incremental change. "Technology will not solve an existing business process problem, and implementing mobile solutions should not be just a costly excuse to implement new technology" (Archer, 2004, p. 180).

FUTURE TRENDS

Previous research addressed the issue of possible human, technical, and financial issues *as well as opportunities and barriers for the possible deployment of* mobile information technology addressing outpatient adherence only at a general level (Cocosila & Archer, 2005b). Future research should look at specific chronic diseases and conditions with high medical, personal, and social impact (e.g., diabetes, asthma, hypertension, congestive heart failure). Quantifiable and nonquantifiable effects of such interventions should be considered when comparing the effects of newer technology with the existing level of care through business lenses.

CONCLUSION

The low level of adherence, especially of chronically ill people in outpatient conditions, is an increasing concern for society because of its consequences for all stakeholders. Novel mobile health care solutions embedded in clinical initiatives may contribute to addressing some of these adherence problems. However, the success of deploying such solutions is a totally new area that must be approached cautiously with a carefully designed, step-by-step, process. This article has raised some questions and offered some suggestions and a step-by-step process framework for developing mobile solutions in health care. Our view is that the real benefit of mobile health care solutions will be realized only when technology and process are built around a logical framework that exploits the particular features of mobile communications and data services.

REFERENCES

Archer, N. (2004). *The business case for employee mobility support.* Paper presented at the International Association for Development of the Information Society (IADIS) (pp. 173-182). Lisbon, Portugal.

Balas, E., & Iakovidis, I. (1999). Distance technologies for patient monitoring. *British Medical Journal, 319*, 1309-1309.

Bashshur, R., Shannon, G., & Sapci, H. (2005). Telemedicine evaluation. *Telemedicine and E-Health, 11*(3), 296-316.

Bludau, H.-B. (2003). *European contribution to mobile computing in medicine.* Retrieved January 30, 2008, from http://www.bbriefings.com/pdf/13/hosp031_r_bludau.PDF

Chin, T. (2002). *Remote control: The growth of home monitoring.* AMNews. Retrieved January 30, 2008, from http://www.ama-assn.org/amednews/2002/11/18/bisa1118.htm

Coambs, R. (2002). The economics of compliance: Managing compliance is one way to contain climbing health care costs. *Canadian Health care Manager, 9*(3), 24.

Cocosila, M., & Archer, N. (2005a). A framework for mobile health care answers to chronically ill outpatient non-adherence. *Informatics in Primary Care, 13*(2), 145-152.

Cocosila, M., & Archer, N. (2005b). Mobile health care initiatives for improving outpatient adherence: Opportunities and barriers. *International Journal of Electronic Business, 3*(6), 512-533.

Cocosila, M., Coursaris, C., & Yuan, Y. (2004). M-health care for patient self-management: A case for diabetics. *International Journal of Electronic Health care, 1*(2), 221-241.

Conlin, R. (2000). *Survey prescribes more net use for doctors*. CRM Daily Website. Retrieved January 30, 2008, from http://www.newsfactor.com/perl/story/5707.html

Cowling, A., Newman, K., & Leigh, S. (1999). Developing a competency framework to support training in evidence-based health care. *International Journal of Health Care Quality Assurance, 12*(4), 149.

Dezii, C. (2000). Medication noncompliance: What is the problem? *Managed Care/Supplement, 9*(9), 7-12.

Duan, M. (2003). *Health care on the wireless frontier.* WirelessInnovator Web site. Retrieved January 30, 2008, from http://cooltown.hp.com/mpulse/1201-health care.asp

Dunbar, P., Madigan, D., Grohskopf, L., Revere, D., Woodward, J., Minstrell, J., et al. (2003). A two-way messaging system to enhance antiretroviral adherence. *Journal of the American Medical Informatics Association, 10*(1), 11-15.

Friedman, R., Stollerman, J., Mahoney, D., & Rozenblyum, L. (1997). The virtual visit. Using telecommunications technology to take care of patients. *Journal of American Medical Informatics Association, 4*(6), 413-425.

Haynes, R. B., McDonald, H. P., & Garg, A. X. (2002). Helping patients follow prescribed treatment: Clinical applications. *Journal of the American Medical Association, 288*, 2880-2883.

Haynes, R., Yao, X., Degani, A., Kripalani, S., Garg, A., & McDonald, H. (2005). Interventions to enhance medication adherence. *The Cochrane Database of Systematic Reviews, 4,* CD000011.

Hoise. (1999). *The newly emerging concept of next generation wireless and mobile telemedicine systems.* The Hoise Website. Retrieved January 30, 2008, from http://www.hoise.com/vmw/99/articles/vmw/LV-VM-05-99-16.htm

Husemann, D. (2004). *Remote monitoring of health conditions.* ERCIM News No. 56. Retrieved January 30, 2008, from http://www.ercim.org/publication/Ercim_News/enw56/husemann.html

Jacobs, D. (1999). Candid comments on wireless. *Health Management Technology, 20*, 8-10.

Junglas, I., & Watson, R. (2003, December 14-17). *U-commerce: An experimental investigation of ubiquity and uniqueness.* Paper presented at the Twenty Fourth Annual International Conference on Information Systems, Seattle, Washington.

Luarn, P., Lin, T., & Lo, P. (2003). An exploratory study of advancing mobilization in the life insurance industry: The case of Taiwan's Nan Shan Life Insurance Corporation. *Internet Research: Electronic Networking Applications and Policy, 13*(4), 297-310.

Maglaveras, N., Koutkias, V., Chouvarda, I., Goulis, D. G., Avramides, A., Adamidis, D., et al. (2002). Home care delivery through the mobile telecommunications platform: The Citizen Health System (CHS) perspective. *International Journal of Medical Informatics, 68*(1-3), 99-111.

McDonald, H. P., Garg, A. X., & Haynes, R. B. (2002). Interventions to enhance patient adherence to medication prescriptions. Scientific review. *Journal of the American Medical Association, 288*, 2868-2879.

Medtronic. (2004). *Medtronic CareLink Network.* Medtronic Website. Retrieved January 30, 2008, from http://www.medtronic.com/carelink/patient/works.html

Mennecke, B., & Strader, T. (Eds.). (2003). Mobile commerce. Technology, theory and applications. Hershey, PA; London, UK: Idea Group Publishing.

Montori, V. M., Helgemoe, P. K., Guyatt, G. H., & Dean, D. S. (2004). Telecare for patients with type 1 diabetes and inadequate glycemic control. *Diabetes Care, 27*(5), 1088-1094.

Poropudas, T. (2001, September 18). *Arbonaut develops medical location with VMWS.* Nordic WirelessWatch, Helsinki. Retrieved January 30, 2008, from http://www.mobile.commerce.net/story.php?story_id=682&s=3

ProWellness. (2004). *Diabetes quality management system.* ProWellness. Retrieved January 30, 2008, from http://www.prowellness.com/pw/english/frontpage_en.htm

Siau, K., Southard, P., & Hong, S. (2002). E-health care strategies and implementation. *International Journal of Health care Technology and Management, 4*(1/2), 118-131.

Tarasewich, P. (2003). Designing mobile commerce applications. *Communications of the ACM, 46*(12), 57-60.

Turisco, F., & Case, J. (2001). *Wireless and mobile computing.* First Consulting Group. Retrieved January 30, 2008, from http://www.chcf.org/topics/view.cfm?itemID=12726

Van Impe, M. (2001, October 15). *Wireless diagnosis funded in the U.K.* Nordic Wireless Watch. Retrieved January 30, 2008, from http://www.nordicwirelesswatch.com/wireless/story.html?story_id=779

Venkatesh, V., Ramesh, V., & Massey, A. (2003). Understanding usability in mobile commerce. *Communications of the ACM, 46*(12), 53-56.

Weiner, J. (2004). *Focusing on side effects increases patient adherence to medication.* Medicine and Behavior. Retrieved January 30, 2008, from http://www.medicineandbehavior.com/mb990813.html

Weingarten, S., Henning, J., Badamgarav, E., Knight, K., Hasselblad, V., JrA, G., et al. (2002, October 26).

Interventions used in disease management programmes for patients with chronic illness-which ones work? Meta-analysis of published reports. *British Medical Journal, 325*, 925.

Wexler, J. (2001). *Wireless apps for the health care industry.* Wireless.itworld.com. Retrieved January 30, 2008, from http://wireless.itworld.com/4246/NWW00322178/pfindex.html

WHO. (2003). *Adherence to long-term therapies: Evidence for action.* World Health Organization Website. Retrieved January 30, 2008, from http://www.who.int/chronic_conditions/adherencereport/en/; http://www.webcitation.org/5FkqlCAQv

Wirelessnewsfactor. (2000). *AT&T to provide wireless health support.* Retrieved January 30, 2008, from www.wirelessnewsfactor.com

Yuan, Y., & Zhang, J. (2003). Towards an appropriate business model for m-commerce. *International Journal of Mobile Communications, 1*(1/2), 35-56.

KEY TERMS

Cost/Benefit Analysis of Health Care Services: The cost of providing/receiving health care services vs. the value of the services, from multiple perspectives.

Disability Adjusted Life Years: Lost years of healthy life caused by premature death or disability.

Mobile Health Care: Using mobile telecommunication technology and portable devices to provide health care services.

Outpatient Conditions: Patients cared for in out-of-hospital conditions (i.e., primary care and home care).

Patient Adherence: A patient's willingness and practice of following medical treatment and advice.

Patient Self-Management: The responsibility and practice of patients managing their own medical care, with the support of advice and education by medical practitioners.

User Adoption: The acceptance and use of a system based on usefulness, ease of use, or other factors.

Managing Operating Room Cost in Cardiac Surgery with Three Process Interventions

Antti Peltokorpi
Helsinki University of Technology, Finland

Juha-Matti Lehtonen
Helsinki University of Technology, Finland

Jaakko Kujala
Helsinki University of Technology, Finland

Juhani Kouri
Kuopio University Hospital, Finland

INTRODUCTION

Health care providers in both public and private sectors are facing increasing pressure to improve their cost efficiency and productivity. The increasing cost of new technological solutions has enforced the application of operations management techniques developed for industrial and service processes. Meyer's (2004) review of existing research shows that, on average, operating rooms (ORs) operate only at 68% capacity. Using OR time efficiently is especially challenging when long operations are scheduled to fixed OR block time. This situation is typical in open heart surgeries where a high variability in the length of required OR time combined with four and a half hour average OR time duration makes scheduling two operations during a normal eight-hour workday difficult.

The objective of this chapter is to analyze the effect that three process interventions have on the OR cost in ORs performing open heart surgeries. The investigated process interventions are four days OR week (4D), the better accuracy of operating room time forecast (F), and doing anesthesia induction outside the OR (I). These interventions emerged from practical organization context.

This chapter is organized as follows. First we provide a review of the existing literature on measures of OR utilization and the investigated three interventions. Based on existing literature, we construct a simulation model to test the interventions' effects on OR utilization. Conclusions of results are presented, and practical implications and new contributions to existing theory of OR management are discussed.

BACKGROUND

Operating room efficiency is typically defined as raw utilization, which means a percentage of time that patients are in the operating room during resource hours (Donham, Mazzei & Jones, 1996). This definition for OR efficiency, however, does not take into account the cost of overused time, which emerges when operations are stretched. Thus, a more valid measure for OR efficiency is a weighted sum of underused and overused OR time (Dexter, 2003). Estimates for relative cost of overused to underused OR time varies in literature from 1.75 (Dexter, Traub & Macario, 2003) to 4 (Dexter, Yue & Dow, 2006). Besides this relative cost, the total OR cost depends on substitutive tasks for underused OR time. Therefore, when evaluating the effect of various process improvements to OR cost, results have to be calculated with case-specific, relative cost for operating time, underused time, and overused time. In the next section, we consider the estimated interventions from existing literature's point of view.

4 Days OR Week (4D)

Especially in long operations, the lengthening of the OR shift has been mentioned as a relevant method to improve OR utilization. In open heart surgeries, the hypothetical advance will be achieved by reallocating the working hours of Friday to the other four working days. Nearly 10 hours OR block time improves the possibility to schedule two operations per day. Despite statistical advances to lengthen the OR block, it has been shown that moving to a four-day week might lead to an increase in sick leaves (Dzoljic et al., 2003). In this case, it was assumed, however, that allocating resources for four 10-hour OR sessions per week is more effective than continuing with 8-hour blocks in open heart surgeries.

The Better Accuracy of Operating Room Time Forecast (F)

An accurate modeling of time distribution as well as an estimation of operating room times has been recognized as an important factor in effective OR planning and scheduling systems (May, Strum & Vargas, 2000; Spangler, Strum, Vargas & May, 2004). The more accurately the required OR time for planned operation can be estimated, the more exactly can the OR resource hours be scheduled while still avoiding overtime. Many researches show that forecasting a required OR time for planned operation is difficult (Macario & Dexter, 1999), but the best estimates can be derived from historical surgeon and surgery type-specific durations (Strum, Sampson, May & Vargas, 2000). Determination of individualized operating room times associated with definition of a daily limit in scheduled OR resource hours resulted in reduction of time overruns and delays before surgery (Broka, Jamart & Louagie, 2003). In open heart surgeries, a hypothetical advance of better OR time forecasting was to get more OR sessions with two operations being performed.

Doing Anesthesia Induction Outside the OR (I)

By moving the anesthesia induction outside the OR and doing it simultaneously with a previous operation, an additional operation can be scheduled for fixed OR block time (Friedman, Sokal, Chang & Berger, 2006). The effect of doing anesthesia induction in parallel with a previous case is best when it makes it possible to increase the amount of performed operations from one to two per day. This is a recurring situation in open heart surgeries. On the other hand, parallel processing requires more nursing staff per OR, and that has to be taken account when assessing a total effect of the intervention (Torkki, Alho, Peltokorpi, Torkki & Kallio, 2006).

RESEARCH QUESTIONS AND METHODOLOGY

The research questions of this chapter are:

1. What is the effect of a four-day OR week, the better accuracy of operating room time forecast, and doing anesthesia induction outside the OR, as well as their combinations, on OR cost per patient?
2. How does the relative cost for operating time, slack time, and overtime impact the effects of the interventions?

The effects of three process interventions and their combinations are tested with a discrete-event simulation model on the open heart surgery patient process. Discrete-event simulation enables the evaluation of alternative productivity improvement proposals while maintaining the dynamic nature of the open heart patient queue. The simulation model was constructed based on Kuopio University Hospital (KUH) open heart surgery process.

CASE UNIVERSITY HOSPITAL

Kuopio University Hospital (KUH) has the special responsibility of treating severe illnesses that call for special expertise and technology in the Hospital District of Northern Savo, Finland. The average yearly number of open heart surgeries is about 1,000. One of the focus areas in KUH is to increase productivity in the operation theater, which was identified as one of the resources that limited the throughput of open heart surgeries. KUH offers coronary bypass surgeries for external hospital districts, and it is losing patients to its competitors. Thus, improving productivity of coronary bypass surgeries would directly increase hospital turnover and profitability.

The input data for developing both the operation time forecasting model and the simulation model was the actual data of all the 2,603 open heart surgeries performed at KUH hospital during the years 2001–2003.

Open Heart Operation Time Model

The average operating theater time usage for open heart surgery is more than four hours, with the actual operation taking 3:21 hours (Table 1). In the simulation model, the OR time is randomly assigned from the historical patient data. A more accurate forecast for needed OR time can be determined based on operation, patient, and surgeon variables (Lehtonen, Kujala, Kouri & Hippaläinen, 2006).

Simulation Model for Evaluation of Production Improvement Proposals

A discrete event simulation model was created to capture the most important elements of the operation theater scheduling system in the case organization (Figure 1).

The emergency patient weekly arrivals are modeled by a Poisson (7,15) distribution. Each of these arrivals is randomly assigned to a weekday. When the assigned weekday is reached, the arriving emergency patient is placed in either an operating room or an emergency patient queue. The OR time as well as OR time forecast are assigned by randomly selecting from the historical emergency patient data. Ten operating rooms per week are reserved for the emergency patients, leaving on average 2.85 rooms per week for elective patients. The emergency surgery overcapacity is supplemented by a specific buffer (fillbuffer), which models the possibility of calling patients from the elective queue at short notice. For each emergency operation, the buffer is checked for the possibility of a second operation during the day. The sum of forecasted operating room times must be less than the length of a workday and of the planned slack. The buffer is filled weekly from the elective queue.

Elective patient arrivals are modeled so they fill the elective queue up to the maximum queue limit. It is assumed that the available demand of operations is larger than operating capacity. Operating time, OR time forecast, and priority are randomized through selecting from historical patient data. There are three priority classes. Priority controls the time a patient may wait in the queue; when the maximum queuing time is reached, the patient is then moved to the front of the queue. Operation planning is done each week for the two operating rooms according to patient priority. For each scheduled patient, a check is preformed to see whether there is another patient in the elective queue whose surgery also can be preformed that day. The OR time that is forecast for the second operation must be less than a day's work in length with the planned slack. If there is no such candidate, only one surgery is scheduled.

To calculate the average cost of surgery for various scenarios, we need to define the relative cost for overtime, slack time, and operating time. Operating time and slack time equals regular hours. Slack time cost depends on how flexibly an organization can reallocate its resources to tasks other than operating during the slack time, such as administration, consultation, or research. The overtime represents the additional cost for overtime work. The following weights for hourly OR time costs were used in the simulation (see Table 2).

The average cost of operation was measured based on the following formula:

$$Cost\_weighted\_OR\_time\_(hr/patient) = \frac{w \times Operating\_time + s \times Slack\_time + o \times Overtime}{Number\_of\_operations}$$

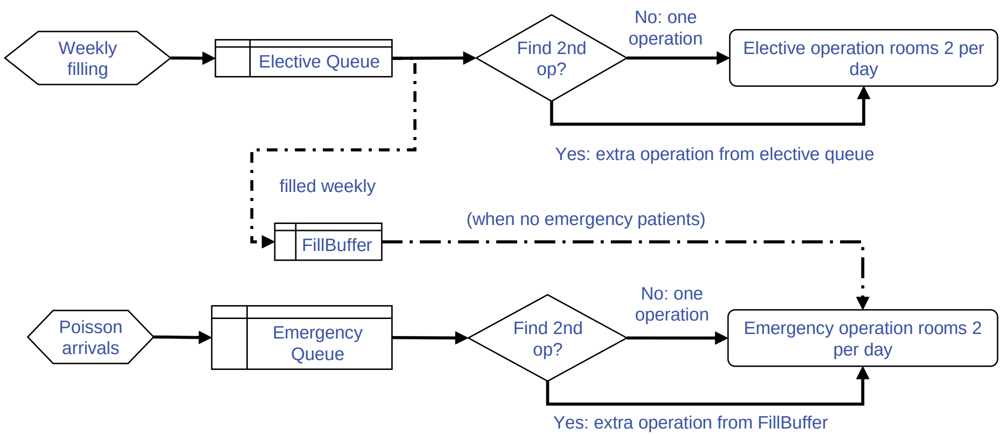
By multiplying the cost weighted OR time by cost of OR hour for the team, we get the average cost per patient (Box 1).

Table 1. The average operating room time usage for open heart surgery (n=2603)

Stage	Start	Room ready	Anaesthesia start	Operation	Anaesthesia end	End	Total
Duration h:mm	0:01	0:24	0:27	3:21	0:16	0:00	4:31
Share %	0.5 %	9.0 %	10.1 %	74.3 %	6.1 %	0.0 %	100.0 %
Standard deviation h:mm	0:03	0:21	0:22	1:21	0:35	0:01	1:22

M

Figure 1. The elements of the KUH operating room scheduling system used in simulation model



Box 1.

$$Average\_cost\_(€/patient) = \frac{w \times Operating\_time + s \times Slack\_time + o \times Overtime}{Number\_of\_operations} \times Cost\_of\_OR\_hour$$

Table 2. Relative cost for operating time, slack time, and overtime used in simulation model

Operating time (w)	Slack time (s)	Overtime (o)
1	1	2
1	1	3
1	1	5
1	0.5	2
1	0.5	3
1	0.5	5

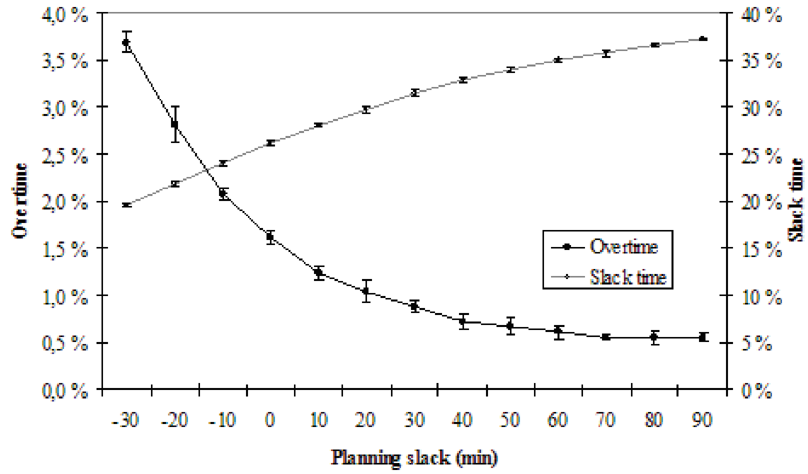
Experimental Variables and Design

The experimental variables included in the simulation experiments were forecast accuracy, induction, and four-day week. The following contains an explanation of how these variables were operationalized in the simulation model experiments:

- **A four-day week (4D):** A change where one of the two elective patient operating rooms is having for a four-day-week instead of a five-day-week, as not everyone is willing to have a four-day week.
- **Forecast accuracy of the OR time (F):** A low value (0) is the forecast based on linear regression, while a high value (1) is a hypothetical improved forecast. The improved forecast is obtained by 0.75 * linear regression forecast + 0.25 * true value.
- **The induction of anesthesia is done outside the operating room in order to reduce other than operation time between two consecutive operations (I):** The reduction achieved by induction when scheduling two operations is around 30 minutes per patient.

The planning slack is the time margin applied between a sum of forecasted OR times and workday length when making a decision of accepting a second opera-

Figure 2. Overtime and slack time with different planning slacks in percentage of cost of regular hours in the "base case" (4D=0, F=0 and I=0), and with slack time penalty s=0.5 and overtime penalty o=5



tion for the day. The effects of the three experimental variables are evaluated over planning slack range from –30 minutes to 90 minutes at 10-minute intervals.

In the simulation experiments, a warm-up period of 100 days was used. Each individual result is based on a three-year run (1,095 days) with 10 replications, without including the warm-up period.

Simulation Results

The chart in Figure 2 shows the overtime and slack time variables in percent of cost of regular hours in the base case (4D=0, F=0 and I=0). The overtime increases at increasing speed as the planned slack is reduced, while the slack time does not decrease as fast as the overtime increases. Also, one can see that there is much more slack time than there is overtime throughout the range of planning slack, which demonstrates the problematic length of open heart surgery in comparison to average working day length.

As the planning slack is decreased, more operations can be performed. This is shown in Figure 2 indirectly, as the slack time decreases and overtime increases as planning slack is reduced.

Table 3 shows for each of the three factors the decrease in costs per patient across all scenarios, each evaluated at the planning slack value that produced optimum results; that is, the average difference between

Table 3. The effect of implementing each of the three interventions in all scenarios at optimum planning slack values. The effect is shown in decreased cost weighted hour per patient and decreased OR cost per patient (calculated with OR time per hour cost for the team = 259 €).

Relative cost		Factor decrease in cost weighted OR time (hr/patient)			Factor decrease in €/patient		
Slack time	Overtime	4-day	Forecast	Induction	4-day	Forecast	Induction
1	2	0.022	0.047	0.407	5.6 €	12.1 €	105 €
	3	0.020	0.068	0.408	5.3 €	17.5 €	106 €
	5	0.027	0.112	0.431	6.9 €	29.0 €	112 €
0.5	2	0.012	0.037	0.327	3.2 €	9.6 €	84.6 €
	3	0.016	0.057	0.330	4.2 €	14.8 €	85.5 €
	5	0.031	0.099	0.325	7.9 €	25.7 €	84.0 €

a factor's highest and lowest values over all the other factor values. The planning slack resulting to optimum value varied between -30 min and 10 min. High overtime valuation, low slack-time valuation, and presence of induction all increased the planning slack value that resulted to optimum.

Factor decrease in euros are calculated by multiplying the cost weighted OR time by 259 €, which is based on OR team staffing (3 doctors and 3.5 nurses) and hourly cost in Finnish wage level (Statistics Finland, 2006, The Finnish Medical Association, 2006). One can see that at the overall level, when using optimal planning slack in scheduling, the most effective intervention is induction. The forecast accuracy has a smaller overall effect but is strongly dependent on weight that is used for overtime work. The four-day week has the smallest impact and appears to be influenced by overtime penalty when s = 0.5 but not that much when s = 1. However, one must remember that the 4D affects only one of the four operating rooms, while forecasting accuracy and induction affect all four. Even in that one room that it affects, is effect clearly smaller than that of induction and, depending on relative costs, smaller or bigger than that of forecasting accuracy?

DISCUSSION

The cost effect of various process improvement interventions depends on relative cost of normal working time, slack time, and overtime. An organization with flexible working hours and an ability to allocate resources to other productive activities for both scheduled and unscheduled slack time is most cost-efficient for performing surgeries whose length is difficult to forecast in advance.

Although performing induction outside the OR seems to be the most efficient intervention in this research, it has to be taken into account that induction typically requires additional personnel (Torkki, Alho, Peltokorpi, Torkki & Kallio, 2006). The actual need for additional personnel, however, depends on type of surgeries and how induction is organized. For example, is it possible for an anesthesiologist to perform anesthesia induction parallel with a previous operation, and are additional nurses required for induction outside the OR?

Compared to existing literature (Friedman et al., 2006; Torkki et al., 2006), this chapter strengthens the argument that performing induction outside the OR is very beneficial when it is possible to add operations to the end of the day. Practically, implementing induction outside the OR requires space and equipment investments.

A four-day week has been mentioned earlier as unwelcome for OR personnel (Dzoljic et al., 2003), and this research also indicates that it does not lead to significant improvements in cost-efficien**cy**. Based on this case, an implementation of a four-day week will need strong management with trade unions.

Based on the literature, improving OR time forecast accuracy is hard work (Macario & Dexter, 1999). This chapter shows that minor improvement in forecasting does not mean remarkable savings in OR cost per patient. Implementing a better forecasting system will require a specific project, and its results are still unpredictable.

FUTURE TRENDS

Recent studies highlight the role of process management methods in operating units. Flexible process and space solutions (Friedman et al., 2006), effective scheduling (Spangler et al., 2004) and personnel incentives are also becoming general in cardiac surgery units. Due to the complexity and multidimensionality of an operating unit, simulation and other computer-based tools will increasingly be utilized when evaluating organizational and process changes.

CONCLUSION

This chapter shows that using anesthesia induction outside the OR improves cost-efficiency. A four-day OR week and better forecasting accuracy have both a positive but relatively smaller impact on OR cost per patient. Forecast accuracy becomes more crucial when the penalty for overtime increases. The benefit of implementing a four-day week and induction are not as much depending on weights for overtime. Based on results, it also can be seen that the effect of all three interventions is more limited if an organization can reallocate its resources to other tasks during the slack time.

For health care managers, this study implicates that the optimal case-specific planning slack times for scheduling can be determined based on discrete event simulation model and by doing anesthesia induction outside the OR in order to reduce OR cost per patient.

In this chapter, a simulation model was used for scheduling patients for operations. The scheduling algorithm as such was not a subject of study (Dexter, Macario & Rodney, 1999). The main benefit of using simulation was its flexibility; that is, the relative ease of developing a descriptive model and the ability of experimenting with it. Simulation can be used in operative scheduling (Lehtonen, Ruohola, Appelqvist & Mattila, 2003), and the simulation model of this article can conceivably also be implemented for OR scheduling, although not without modifications and linking it to existing information systems.

REFERENCES

Broka, S., Jamart, J., & Louagie Y. (2003). Scheduling of elective surgical cases within allocated block-times: Can the future be drawn from the experience of the past? *Acta Chirurgica Belgica, 103*, 90–94.

Dexter, F. (2003). Operating room utilization: Information management systems. *Current Opinion in Anaesthesiology, 16*, 619–622.

Dexter, F., Macario, A., & Rodney, T. (1999). Which algorithm for scheduling add-on elective cases maximizes operating room utilization? *Anesthesiology, 91*, 1491–500.

Dexter, F., Traub, R., & Macario, A. (2003). How to release allocated operating room time to increase efficiency—Predicting which surgical service will have the most under-utilized operating room time. *Anesthesia & Analgesia, 96*, 507–512.

Dexter, F., Yue, J., & Dow, A. (2006). Predicting anesthesia times for diagnostic and interventional radiological procedures. *Anesthesia & Analgesia, 102*, 1491–1500.

Donham, R., Mazzei, W., & Jones, R. (1996). Glossary of times used for scheduling and monitoring of diagnostic and therapeutic procedures. *American Journal of Anesthesia and Analgesia, 97*, 1127–1132.

Dzoljik, M., Zimmerman, M., Legemate, D., & L;azomga. N. (2003). Reduced nurse working time and surgical productivity and economics. *Anesthesia & Analgesia, 97*, 1127-1132.

Finnish Medical Association. (2006). Lääkäreiden palkkatietoa (in Finnish). Retrieved June 19, 2006, from http://www.tilastokeskus.fi/til/vkp/index.html.

Friedman, D., Sokal, S., Chang, Y., & Berger, D. (2006). Increasing operating room efficiency through parallel processing. *Annals of Surgery, 243*(1), 10–14.

Lehtonen, J-M., Kujala, J., Kouri, J., & Hippeläinen, M. (2006). Productivity and throughput improvement in cardiac surgery [submitted for review]. *International Journal for Health Care Quality Assurance*.

Lehtonen, J-M., Ruohola, T., Appelqvist, P., & Mattila, I. (2003). Simulation-based finite scheduling at Albany International. In *Proceedings of the Winter Simulation Conference (WSC'03)*, New Orleans, Louisiana.

Macario, A., & Dexter, F. (1999). Estimating the duration of a case when the surgeon has not recently scheduled the procedure at the surgical suite. *Anesthesia and Analgesia, 89*, 1241–1245.

May, J., Strum, D., & Vargas, L. (2000). Fitting the lognormal distribution to surgical procedure times. *Decision Sciences*, *31*, 129–148.

Meyer, M. (2004). Perioperative surgery in the twenty-first century—Two case studies. *AORN Journal*.

Spangler, W., Strum, D., Vargas, L., & May, J. (2004). Estimating procedure times for surgeries by determining location parameters for the lognormal model. *Health Care Management Science*, 7, 97–104.

Statistics Finland. (2006). Central government monthly wages and salaries. Retrieved June 19, 2006, from http://www.tilastokeskus.fi/til/vkp/index.html

Strum, D., Sampson, A., May, J., & Vargas, L. (2000). Surgeon and type of anesthesia predict variability in surgical procedure times. *Anesthesiology, 92*, 1454–1167.

Torkki, P., Alho, A., Peltokorpi, A., Torkki, M., & Kallio, P. (2006). Managing urgent surgery as a process—Case study of a trauma center. *International Journal of Technology Assessment in Health Care, 22* (2), 255–260.

KEY TERMS

Anesthesia Induction: Time when the anesthesiologist begins the administration of agents intended to provide the level of anesthesia required for the scheduled procedure.

Discrete Event Simulation: A simulation model where significant changes occur at discrete time instances.

Operating Time: Time patient is in the operating room during regular hours.

Overtime: Time that patient is in the operating room after regular hours.

Planning Slack Time: Minimum time that has to be left nonscheduled at the end of regular hours.

Slack Time: Time operating room is empty at the end of regular hours.

Managing Paramedic Knowledge for Treatment of Acute Myocardial Infarction

Tom Quinn
Coventry University, UK

R. K. Bali
Coventry University, UK

Kathrine Dale
Worcestershire Royal Hospital, UK

Pete Gregory
Coventry University, UK

INTRODUCTION

There are about 280,000 acute myocardial infarctions (AMI) each year in the United Kingdom (British Heart Foundation (BHF), 2004). Up to half of these events will prove to be fatal, many within the first few hours following symptom onset as a result of disturbances of cardiac rhythm. For those patients who survive the initial period of vulnerability, subsequent mortality and the risk of disabling chronic heart failure can be markedly reduced by prompt restoration of blood flow through the affected coronary artery.

This can be achieved in patients sustaining a STEMI by mechanical means (percutaneous coronary intervention) or pharmacologically by thrombolytic ("clot buster") therapy (Antman et al., 2004) or where initial thrombolytic treatment fails—"rescue" intervention. Mortality benefit from thrombolysis is time dependent; for patients presenting with STEMI in the first 2 to 3 hours from onset, each minute's delay equates to 11 days loss of life expectancy (Department of Health (DoH), 2003).

BACKGROUND

In England, publication of the National Service Framework (NSF) for coronary heart disease (CHD) has resulted in dramatic improvements in the timeliness of thrombolytic treatment (DoH, 2000), with half of eligible STEMI patients starting treatment within the NSF standard of 60 minutes from the call for help (Birkhead et al., 2004) compared to one in 10 in the years preceding the NSF (Quinn, Allan, Birkhead, Griffiths, Gyde & Murray, 2003). These improvements are largely a result of improved hospital processes (Birkhead et al., 2004), but further increasing the proportion of patients treated within the NSF "call to needle time" is reliant on thrombolysis commencing prior to hospitalization (Quinn et al., 2003), administered mainly by ambulance paramedics.

Significant government investment in emergency ambulance services in England has facilitated universal availability of 12 lead electrocardiograph (ECG) equipment, programs to enhance training of paramedics in assessment of patients with suspected STEMI for thrombolysis eligibility (and other aspects of cardiac care), and procurement of the appropriate thrombolytic drugs.

The United Kingdom provides a mainly paramedic-led ambulance service in contrast to several other countries where some ambulances are manned by physicians or intensive-care trained nurses (Bouten, Simoons, Hartman, Van Miltenburg, Van der Does & Pool, 1992; Leizorovicz et al., 1997). This has led to concern in some quarters about the feasibility and safety of nonphysicians administering thrombolysis in the ambulance setting.

Recent UK studies, however, have confirmed that paramedics can accurately identify patients with suspected AMI requiring specialist coronary care (Quinn, Allan, Thompson, Pawelec & Boyle, 1999) and, in particular, patients presenting with ST segment elevation (the cornerstone of decision-making in this context) on the ECG (Whitbread et al., 2002) and eligibility for thrombolysis (Pitt, 2002; Keeling et al., 2003). Moreover, a recent international clinical trial has demonstrated that the efficacy and safety of prehospital thrombolysis was independent of physician presence (Welsh et al., 2005). Around 4,000 patients in England have received "paramedic thrombolysis" since 2003.

Nevertheless, the government's own Review of Early Thrombolysis (DoH, 2003) has acknowledged concerns about paramedics' retention of knowledge and skill in relation to thrombolytic treatment; in particular, in relation to the low exposure to this aspect of emergency care in low-density populations in rural areas. The review acknowledges the need for research to address the question of a "volume-outcome" relationship between paramedics' exposure to the procedure and patient safety.

LEARNING, SKILL, AND KNOWLEDGE RETENTION

It is appropriate to make a distinction between learning (a change in behavior as a result of experience) (Haskell, 2001) and skill retention (persistence of skill proficiency after a period of no practice) (Cree & Macauley, 2003).

Although there is a paucity of research specific to the areas of learning and skill retention around paramedics and thrombolytic treatment, the literature on other aspects of paramedic practice may be relevant. For example, paramedics' skill acquisition, retention, and decay have been assessed in relation to tracheal intubation (Barnes, 2001), neonatal resuscitation (Skidmore & Urquhart, 2001), and laryngeal mask intubation (Coles, Elding & Mercer, 2001). Studies of skill decay in the delivery of cardiopulmonary resuscitation demonstrate poor skill and knowledge retention with a return to pretraining levels within six months (Roach & Medina, 1994). Issues of motivation, competence, confidence, and assessment come in to play (Chamberlain & Hazinski, 2003).

COMPETENCE

While these are essentially psychomotor skills, the safe administration of thrombolytic treatment requires paramedics to demonstrate a range of competencies (Quinn, Butters & Todd, 2002) involving patient assessment, ECG interpretation, identification of contraindications to treatment (the latter largely aimed at reducing risk of intracranial and other haemorrhagic complications), weight estimation (Hall et al., 2004), and explaining the risks and benefits of treatment to patients.

The presence of a so-called rural paramedic paradox has been suggested in which patients living further away from hospitals tend to be in low-density, rural populations where an individual paramedic's exposure to severe emergencies (e.g., STEMI) may be low, but the distance and time to reach a hospital may mean that the need for highly skilled paramedic intervention is highest (Rowley, 2001). Whether this results in poorer patient care is unknown.

A volume outcome relationship for STEMI patients has been reported in relationship to physician exposure to the condition (Tu, Austin & Chan, 2001). However, with an estimated 700,000 emergency attendances with chest pain to hospitals in England and Wales per annum (Goodacre et al., 2005), it is likely that paramedics will be performing many more assessments and ECG recordings than they will be administering thrombolysis to or referring for primary intervention; therefore, the key assessment skills will be subject to frequent exposure.

KNOWLEDGE MANAGEMENT

A number of studies have put forth the notion of incorporating KM in health care as a way out of the imminent information explosion crisis (Koretz & Lee, 1998; Wyatt, 2001). The average physician spends as much as a quarter of his or her professional activities managing information and has to learn an estimated 2 million clinical specifics (Wyatt, 2000), a burden compounded by the doubling of biomedical literature every two decades. The burden on other health professionals, including paramedics, has not been reported. Paramedics, as registered health professionals, also are required to maintain their knowledge base, although currently didactic clinical guidelines are prepared and distributed nationally. As with other generalist health professionals, paramedics' scope of practice (and thus, underpinning knowledge) is potentially enormous.

"Knowledge is the enemy of disease" (Brice & Gray, 2003) and serves ultimately to protect the patient from harm. Data from NHS information sources (Brice & Gray, 2003) indicate the scale of the problem of serious failures in health care. From these sources, we learn that annually:

- 400 people die or are seriously injured in adverse events involving medical devices.
- Nearly 10,000 people are reported to have experienced serious adverse reactions to drugs.
- Around 1,150 people who have been in recent contact with mental health services commit suicide.
- Nearly 28,000 written complaints are made about aspects of clinical treatment in hospitals.
- The NHS pays out around £400 million a year for settlement of clinical negligence claims, and has a potential liability of around £2.4 billion for existing and expected claims.
- Hospital acquired infections, around 15% of which may be avoidable, are estimated to cost the NHS nearly £1 billion.

As skills and knowledge exist in both individuals and organizations, the essence of knowledge management is how best to identify and capture knowledge and experience in order to disseminate it further throughout the organization. There have been several attempts to bring various forms of clinical information to the clinician at the point of care. They include:

- Developing computer applications that stand alone, although they are often network accessible and are available to the clinician upon his or her request.
- Incorporating the clinical knowledge directly into clinical information systems used by clinicians while giving care (once there, the computer-based system can automatically prompt the clinician or the clinician can request help).
- Requesting help from an outside source (Sittig, 2005).

FUTURE TRENDS

While many of these approaches relate to physicians in primary and secondary care settings, the use of computerized decision aides in emergency cardiac care has been reported (Grisjeels et al., 1996; Lamfers et al., 2004); although these are not widely used in the ambulance service in England, the development of the National Programme for Information Technology across the National Health Service (NHS) raises new opportunities (Humber, 2004), including wider utilization of on-board computers on emergency ambulances (Bastow, 2003).

Paramedics often are required to make rapid life or death decisions that can have far-reaching consequences for patients. As previously stated, these decisions include such competencies as patient assessment, ECG interpretation, identification of contraindications to treatment, weight estimation, and explaining the risks and benefits of treatment to patients. Therefore, paramedics need to embrace appropriately designed, tested, and reliable tools and techniques of knowledge management to facilitate rapid and safe processing of pertinent knowledge.

IT-based clinical knowledge is managed by way of such tools as medical decision support systems, which themselves are a product of clinical practice guidelines issued by various government and medical associations. Clinical practice guidelines are effectively rule-based knowledge that guides the decision makers in a medical setting (Field & Lohr, 1992). These guidelines have been developed over the years to reduce the variations among medical practices with a common goal to provide cost-effective and high-quality health care services.

This has to be balanced against the reality of the contemporary environment and associated employee performance and safe practice within this rapidly changing practice environment. There is an acknowledged gap between the importance of assessing what employees can do and not what they know. Del Bueno, Weeks, and Brown-Stewart (1987) report a study carried out among test takers who have difficulty performing a procedure or recognizing warning signs in a patient experiencing difficulty. The use of criterion-based performance measures determines practice competencies in employees as well as identifies where need exists to correct skill or knowledge deficiencies.

TOWARD A CARDIAC CARE KNOWLEDGE MANAGEMENT SYSTEM (KMS) FOR PARAMEDICS

Knowledge Management Systems (KMS) consist of appropriate information technologies (IT), policies, processes, and procedures to manage the creation, storing, sharing, and reuse of knowledge. In this context, knowledge is any information that is relevant and actionable and is based on a person's experience (Davenport, De Long & Beers, 1998). For paramedics, an accessible cardiac care KMS would assist in the reduction of health care costs, which are growing at an unsustainable rate (up 20% since 2001) and placing a major burden on the economy (Warner, 2004). KMSs have the potential to reduce medical errors, which cause an estimated 1 million injuries and 98,000 deaths each year (Davenport & Glaser, 2002; Warner, 2004).

For the creation of a viable cardiac care KMS, further work is required to assist health care organizations in carrying out activities better, faster, and cheaper (Dwivedi, Bali, James, Naguib & Johnson, 2002). Such research work would include (a) analyzing and forming a suitable KM strategy for the paramedic function; (b) adopting best practice from established KM strategies and implementations (from the nonclinical environments); and (c) capturing and integrating paramedic perspectives on clinical knowledge which few previous studies have managed effectively.

CONCLUSION

In a publicly funded health system driven by accountability, defining competence has been described as one of the most challenging and essential components in the accountability paradigm (Redman, Lenburg & Hinton Walker, 1999). With more and more medical interventions such as thrombolysis being delegated to nonphysicians (and in this case, particularly paramedics), there is a need for improved understanding of how knowledge is managed and how best to optimize effective clinical decision-making.

REFERENCES

Antman, E., et al. (2004). ACC/AHA guidelines for the management of patients with ST-elevation myocardial infarction: A report of the American College of Cardiology/American Heart Association task force on practice guidelines (committee to revise the 1999 guidelines for the management of patients with acute myocardial infarction). *Journal of the American College of Cardiology, 44,* E1-E211.

Barnes, T.A. (2001). Tracheal intubation: Training and retention. *Respiratory Care Journal, 46*(3).

Bastow, P. (2003). Electronic patient-report forms in the ambulance service: Ready for national rollout. *British Journal of Healthcare Computing Information Management, 20*, 44-46.

Birkhead, J., et al. (2004). Improving care for patients with acute coronary syndromes: Initial results from the National Audit of Myocardial Infarction Project (MINAP). *Heart, 90*, 1004–1009.

Bouten, M.J.M., Simoons, M.L., Hartman, J.A.M., Van Miltenburg, A.J.M., Van der Does, E., & Pool, J. (1992). Prehospital thrombolysis with alteplase (rt-PA) in acute myocardial infarction. *European Heart Journal, 13*, 925–931.

Brice, A., & Gray, M. (2003). Knowledge is the enemy of disease. An organisation with a memory. *Library and Information Update, Department of Health, 2*(3), 32–34.

British Heart Foundation. (2004). Coronary heart disease statistics database. Retrieved February 21, 2005, from www.heartstats.org.

Chamberlain, D.A., & Hazinski, M.F. (2003). Education in resuscitation. *Resuscitation, 59*, 11–43.

Coles, C., Elding, C., & Mercer, M. (2001). A laboratory assessment of the learning and retention skills required to use the Combitube and Laryngeal Mask Airway by non-anaesthetists. *Critical Care, 5*(Suppl), 1 4.

Cree, V., & Macauley, C. (2003). *Transfer of learning in professional and vocational education*. Routledge.

Davenport, T.H., De Long, D.W., & Beers, M.C. (1998). Successful knowledge management projects. *Sloan Management Review*, Winter, 43–57.

Davenport, T.H., & Glaser, J. (2002). Just-in-time delivery comes to knowledge management. *Harvard Business Review*, July, 107–111.

Del Bueno, D.J., Weeks, L., & Brown-Stewart, P. (1987). Clinical assessment centers: A cost-effective alternative for competency development. *Nursing Economics. 5*(1), 21–26.

Department of Health. (2000). *National service framework for coronary heart disease*. London: Department of Health.

Department of Health. (2003). Review of early thrombolysis—Faster and better treatment for heart attack patients. London: Department of Health.

Dwivedi, A., Bali, R.K., James, A.E., Naguib, R.N.G., & Johnson, D. (2002). Towards a holistic knowledge management framework for healthcare institutions. In *Proceedings of the Second Joint EMBS/BMES Conference* (pp. 1894-1895), Houston, TX.

Field, M.J., & Lohr, K.B. (1992). Guidelines for clinical practice: From development to use. Washington, DC: National Academy Press.

Goodacre, S., et al (2005). The health care burden of acute chest pain. *Heart, 91*, 229–230.

Grisjeels, E.W.M., et al. (1996). Implementation of a pre-hospital decision rule in general practice. Triage of patients with suspected myocardial infarction. *European Heart Journal, 17*, 89–95.

Hall, W.L., et al. (2004). Errors in weight estimation in the emergency department: Comparing performance by providers and patients. *Journal of Emergency Medicine, 27*, 219–224.

Haskell, R.E. (2001). *Transfer of learning: Cognition, instruction and reasoning*. San Diego: Academic Press.

Humber, M. (2004). National programme for information technology. *British Medical Journal, 328*, 1145–1146.

Keeling, P., et al. (2003). Safety and feasibility of prehospital thrombolysis carried out by paramedics. *British Medical Journal, 327*, 27–28.

Koretz, S., & Lee, G. (1998). Knowledge management and drug development. *Journal of Knowledge Management, 2*(2), 53–58.

Lamfers, E.J., et al. (2004). Prehospital versus hospital fibrinolytic therapy using automated versus cardiologist electrocardiographic diagnosis of myocardial infarction: Abortion of myocardial infarction and unjustified fibrinolytic therapy. *American Heart Journal, 147*, 509–515.

Leizorovicz, A., et al. (1997). Pre-hospital and in-hospital delays in thrombolytic treatment in patients with suspected acute myocardial infarction. Analysis of data from the EMIP study. European Myocardial Infarction Project. *European Heart Journal, 18*, 248–253.

Pitt, K. (2002). Prehospital selection of patients for thrombolysis by paramedics. *Emergency Medical Journal, 19*, 260–263.

Quinn, T., Allan, T.F., Birkhead, J., Griffiths, R., Gyde, S.N., & Murray, R.G. (2003). Impact of a region-wide approach to improving systems for heart attack care: The West Midlands Thrombolysis Project. *European Journal of Cardiovascular Nursing, 2*, 131–139.

Quinn, T., Allan, T.F., Thompson, D.R., Pawelec, J., & Boyle, R.M. (1999). Identification of patients suitable for direct admission to a coronary care unit by ambulance paramedics. *Pre-hospital Immediate Care, 3*, 126–130.

Quinn, T., Butters, A., & Todd, I. (2002). Implementing paramedic thrombolysis: An overview. *Accident and Emergency Nursing, 10*, 189–196.

Redman, R.W., Lenburg, C.B., & Hinton Walker, P. (1999). Competency assessment: Methods for development and implementation in nursing education. *Journal of Issues in Nursing*. Retrieved September 30, 1999, from www.nursingworld.org/ojin/topic10/tpc10_3.htm.

Roach, C.L., & Medina, F.A. (1994). Paramedic comfort level with children in medical and trauma emergencies: Does the PALS course make a difference. *American Journal of Emergency Medicine, 12*, 260–262.

Rowley, T. (2001). Solving the paramedic paradox. *Rural Health News, 8*, 3.

Sittig, D.F. (2005). An overview of efforts to bring clinical knowledge to the point of care. In R.K. Bali (Ed.), *Clinical knowledge management: Opportunities and challenges*. IGP.

Skidmore, M.B., & Urquhart, H. (2001). Retention of skills in neonatal resuscitation. *Paediatrics and Child Health, 6*(1), 31–35.

Tu, J.V., Austin, P.C., & Chan, B.T. (2001). Relationship between annual volume of patients treated by admitting physician and mortality after acute myocardial infarction. *Journal of the American Medical Association, 285*, 3116–3122.

Warner, M. (2004). Under the knife. *Business 2.0, 5*(1), 84–89.

Welsh, R.C., et al. (2005). Time to treatment and the impact of a physician on pre-hospital management of acute ST elevation myocardial infarction: Insights from the ASSENT–3 PLUS trial. *Heart, 91*(11), 1400–1406.

Whitbread, M., et al. (2002). Recognition of ST elevation by paramedics. *Emergency Medicine Journal, 19*, 66–67.

Wyatt, J.C. (2000). Intranets. *Journal of the Royal Society of Medicine, 93*(10), 530–534.

Wyatt, J.C. (2001). Management of explicit and tacit knowledge. *Journal of the Royal Society of Medicine, 94*(1), 6–9.

KEY TERMS

Emergency Cardiac Care: Includes resuscitation from cardiac arrest and care of patients with suspected heart attack (acute myocardial infarction) and its complications.

Knowledge Management: The collection, organization, analysis, and sharing of information held by workers and groups within an organization.

Myocardial Infarction: Another term for a heart attack; occurs when a blood vessel (an artery) to the heart is blocked.

Paramedic: Registered health care professional who provides emergency and urgent care and treatment to patients who are either acutely ill or injured; paramedics usually work outside the hospital environment and are able to perform a variety of advanced skills, administer a range of drugs, and carry out certain surgical techniques.

Thrombolysis: Use of drugs to break down a blood clot that is obstructing a blood vessel (usually an artery).

Mapping Information of Operating Theatre Waiting List Process

Gerhardine Foo
University of Southern Queensland, Australia

Latif Al-Hakim
University of Southern Queensland, Australia

INTRODUCTION

An operating theatre waiting list (OTWL) is a list that patients are enrolled in once they opt to pursue an elective procedure, assuming they cannot get this procedure performed immediately (Chua, 2005). Operating Theatre Waiting Lists are of great concern in society nowadays because of their societal and political priority, their link to the quality of individual patients' lives, relation to the economic management of operating theatres and management of patient flow through the hospital, and distribution of scarce medical resources (Al-Hakim & Fitzgerald, 2003; Foote, North & Houston, 2004; NSW Health, 2002). They are used by politicians as measures of success or otherwise measures of government action on the health services.

It is generally accepted that waiting lists are unacceptably long and create a number of problems (Foote et al., 2004). Long waiting times are a problem for patients, not only because of uncertainty but also because the state of the patient may deteriorate if he or she is not treated early enough (Council of Europe, 2005). Nothing is more important to patients than the length of time they have to wait for treatments. The impact of fast, effective treatment, diagnosis and treatment on a patient's health and well-being is incalculable (Scottish Executive, 2001). Disruptions and delay affect the quality of patient care and may cause stress, anxiety, and discomfort for patients as well as cause costs to escalate as sessions overrun and much more (Buchanan & Wilson, 1996). Also, if a patient does not arrive in the operating theatre on time, the surgeon may be left idle or left to perform routine duties, which is expensive for the hospital and frustrating for the staff (Buchanan, 1998). There also have been reports about the depressing effects on patients of cancelling surgeries and on the high level of emotional involvement before surgeries. Late cancellation of scheduled operations is also a major cause of inefficient use of operating-room time and a waste of resources; it is potentially stressful and costly to patients in terms of working days lost and disruption to daily life as well (Schofield et al., 2005).

Sometimes waiting lists are so long that patients have to be turned away. In September 2005, it was found that more than 30 patients were turned away daily from the Royal Brisbane and Women's Hospital (Sommerfeld, 2005). Long waiting lists also have been found to distort the clinical judgement of doctors, thereby causing suffering for some of the most seriously ill patients. It has been found that half of the consultants in England admitted deferring surgery in high priority cases to treat less urgent patients because of the pressure to cut waiting lists in those areas (Carvel, 2001). The consultants who admitted to treating patients in the wrong order said that this had a negative impact on the patients' conditions.

OTWL Activities

OTWL comprises a large number of activities that can be grouped into five main categories:

1. **Pre-admission.** This includes activities related to the referral from the GP or specialist and the deciding of the level of categorization that the patient is given as well as the booking of appointment.
2. **Peri-operative.** This includes the activities of patient booking, operating scheduling, and patient preparation. It consists of:
    - Expanding the scope of peri-operative care to include all activities from the hospital receiving the recommendation for admission form to discharge from either the day-only or recovery unit

- An expanded scope of responsibility—bundling activities with responsibility delegated to an individual to enable new processes to occur
- A collocation of related activities—booking and scheduling with the pre-admission clinic, day-only and day-of-surgery unit located adjacent to the operating theatres

3. **Intra-operative.** Involves managing operating theatre resources and procedures directly associated with the surgery to ensure safe and effective patient outcomes.
4. **Bed management.** Concerned with making different types of beds available to patients. Relevant admission and discharge policies are required to manage bed availability on an hour-to-hour, day-to-day basis. Examples identified include elective day surgery admissions and patients cared for in the ambulatory unit following surgery.
5. **Discharge planning.** Encompasses activities aimed at the smooth transition of the patient from hospital to their home or place of residence.
6. **The information flow.** Involves the exchange and use of data to manage surgical services and operating theatres. These include:
   - Surgeon/specialty session allocation, including changes due to individual surgeon's availability throughout the year
   - Surgeon and hospital waiting list data
   - Seasonal/population demands
   - Patient data and clinical information
   - Operating theatre booking and scheduling, including resources, equipment, and prostheses
   - Patient bookings, including appointment dates and times for the pre-admission clinic, and admission dates and times for day-only and day-of-surgery admissions
   - Real time activity data to manage elective and nonelective/emergency cases
   - Availability of acute beds, including intensive care unit and high dependency unit beds and the number of patients being discharged
   - Hospital and operating theatre utilization and outcome data; for example, pre-admission clinic attendance rates, day-only and day-of-surgery rates, and cancellation rates

Potential Causes of the Problem

There is often a lack of consistency between planned theatre session lists and the manner in which these actually run. Theatre staff and users frequently fail to keep to regular session start and finish times; thus, late starts and overruns of sessions are commonplace. This is because once the list has been drawn up, it is still vulnerable to multiple alterations and modifications (Audit Commission, 2002; 2003; Buchanan & Wilson, 1996; Healthcare Commission, 2005; McAleer, Turner, Lismore & Naqvi, 1995; Schofield et al., 2005). Following is a list of the main causes of the problem of long OTWL:

- Accommodating emergency cases that are given priority
- No separate emergency theatre session, leading to interruptions in routine lists
- Over-ambitious, poor list planning
- Unanticipated variation in the patient's condition, only known during the surgery
- Unpredictability of procedure's length of time
- Unanticipated variations in the time of patient's stay in the recovery ward
- Cancellation of operations on the day of intended surgery
- Preoperation tests are not done properly or completed on time
- Unavailability of beds in recovery wards
- No agreed start and finish times for theatre sessions
- Lack of commitment by theatre staff and users to ensure they arrive on time
- Delays due to external factors such as patients, equipment, and supplies not being delivered to the theatre on time
- Patients delaying surgery due to personal reasons (e.g., fear)
- Unavailability of resources, such as specialists, porters, surgical equipment, and so forth

Sometimes a poorly maintained information system also contributes to the problem of the waiting list. Generally, decisions about theatre lists are entered in multiple patient records by a number of staff and officers who do not always communicate their independent actions. This is not helped by the fact that there is difficulty arranging meetings with all the specialists, who are very busy people with overcrowded schedules (McAleer et al., 1995). The reasons and causes identified previously are not exhaustive. They only serve to indicate the complexity of the problem and the importance of developing an effective methodology for mapping information flow of the OTWL process.

Traditional process mapping methodologies cannot recognize the various activities' elements other than their inputs and outputs. Accordingly, these methodologies lack the ability to determine the interdependencies among all elements of process activities and, accordingly, cannot be used to map information flow among all these elements or interdependencies. Failure to map information flow comprehensively may affect the successful implementation of BPR.

The objective of this chapter is to develop a theoretical framework for mapping information between various activities' elements of the OTWL process that provides a valid platform for evaluating various OTWL scenarios to enable the most applicable design to be selected. It is hoped that with a better design to coordinate the activities of the OTWL process, the quality of health care will improve. This research is part of a larger project for redesigning the operating theatre process (Project No. 1000819), which is supported by the Faculty of Business at the University of Southern Queensland. The project is initiated by the second author and comprises two research students: a Master's Degree researcher and an Honour Degree researcher (the first author).

SOLUTION APPROACHES

Researchers have used various ways to attempt to solve the problem of long OTWL; namely, by using various kinds of techniques to identify the problems. Operations research, in particular, has been popular. Operations research is concerned with the conduct and coordination of activities within complex systems, using tools such as mathematical modeling, queuing theory, and simulation to study the consequences of alternative courses of actions and to optimise performances of the system (Buhaug, 2002). The aim of operations research is to provide numerical answers to problems and insights into the nature of the problem as well, thereby helping hospital staff and managers to ask the right questions, make sense of the answers, and look in the right direction for solutions. Some of the operations research approaches are discussed here.

McAleer, et al. (1995) used Visual Interactive Simulation (VIS) for the simulation of a hospital's theatre suite. The study was undertaken to identify delays and restrictions in the system (due to resources imbalance) by using a VIS model to explore the relationship among various throughput levels and resource allocations. However, the aim is to only optimise the use of resources, and they did not take into account the interdependencies of the activities.

The simulation software program Witness was used by Ivaldi, Elena, and Testi (2003) to develop a discrete-event simulation model of a surgical department. Ramirez-Valdevia and Crowe (1997) attempted to use a methodology called Simulation Service Quality System (SSQS) to improve operating performance measures in light of a patient's preferences. Patients' expectations are used to set performance standards, and then a model of such is simulated using the discrete-event simulation package SIMAN.

Most of the research work analyzing waiting lists in Spain has been based on queuing theory. Gonzalez-Busto and Garcia (1999) attempted a systems dynamic approach to solve the waiting list problem in Spanish hospitals. Their research involves building a simulation model in order to improve the understanding of how waiting lists behave over time and to analyze the long-term effectiveness of the policies commonly used when managing them. This simulation model includes the policies currently in practice (subcontracting activities to public and privately owned hospitals, the so-called special programs, waiting list updating or revision) and some new proposals (capacity investment decision and staggering of vacation periods).

Attempts to build Priority Queuing Models to reduce waiting times in emergency care (Johnson, Jones & Siddharthan, 1996) have been attempted as well. Their claim as the solution to the problem is to force nonemergency users to realize the social costs of their actions and to impose additional waiting times.

Foote, North and Houston (2004) report findings of case studies of problematic ultrasound waiting lists and provide some insight into the understanding of hospital waiting lists. They highlighted that waiting lists are better conceptualized as a set of interacting issues instead of viewing them as a rational queue, and that technical solutions would not be sufficient to address the problem.

Barlow (2002) tried to find a critical factor affecting the waiting list by measuring the gap between the expectations and perceptions of the OTWL services of the patient. He concluded that the actual time of delay is not of significant importance; rather, it is the difference between the customer perceptions and expectation that create the level of satisfaction.

Methods such as scheduling have been used but have proved to be too simplistic and insufficient in addressing the complexity of the OTWL process. This can be seen in the NSW Health Operating Theatre Management Project conducted by KPMG consultants (NSW Health, 2002). Despite the huge financial investment in the project, the end result was a simple scheduling and identification of the major activities. There was insufficient description of specific details; for instance, the transfer of a referral letter from a doctor.

Buchanan and Wilson (1996) tried to use Business Process Re-engineering to redesign the surgical patient trail from referral to a specialist by a general practitioner through admission, treatment, and discharge. Motives for doing so include cost reduction, improvements in theatre's capacity utilization, improvements in staff morale, and quality of patient care.

The use of the fuzzy logics approach to scheduling to support decision-making about queues in operating theatre suites has been used (Al-Hakim, 2006; Al-Hakim & Fitzgerald 2003). It is a very deterministic approach in the sense that it gives a yes or no answer instead of more or less.

OTWL PROCESS MAPPING AND INFORMATION MAPPING

A process cuts across the functional boundaries of an organization and is the means by which an organization achieves its stated objectives (Bal, 1998). Process complexity is often difficult to visualize. To adequately address the issues, there is a need to map the OTWL process. Process mapping is a technique used to detail business process activities by focusing on the important elements that influence those activities (Soliman, 1998); it consists of constructing a model that shows the relationships among the activities, people, data, and objects involved in the production of a specified output (Biazzo, 2002). Process mapping is used to re-engineer or improve a system and encourages a process orientation and overview (Buchanan, 1998); its models can offer useful and relatively inexpensive descriptions that can help toward improving and redesigning business processes (Biazzo, 2002). With staff participation, it broadens cross-functional awareness and understanding; potentially fosters mutual respect for different contributions to the patient trail; and can deepen appreciation of the extent to which problems are shared, which activities may be unnecessarily duplicated, and how problems can be unwittingly passed on from one stage on the trail to another (Buchanan, 1998).

To successfully map a process, we need to go beyond the detailing of each activity in the process. There is a need to understand the interdependencies among various elements of the process as well as the flow of information. In other words, there is a need to adequately map OTWL process (Al-Hakim, 2006). To do so, the first step is to comprehend what constitutes an activity—the elements of an activity.

Generally, OTWL comprises two main aspects: technical and social. Therefore, sociotechnical theory is applicable here in addressing these issues (Fernandez, Jacobs, Kauffman & Keating, 2001). The previous studies of OTWL mainly examined the technical aspects of the process with a few mentioning but not taking into account social factors, or vice versa.

Most of the typical process mapping efforts such as flowcharts and Gantt charts show only how the input and output of activities are linked. The direct consequence of conceptualizing processes as flowcharts is an implicit assumption that the problem of process innovation can be faced using the hard engineering approach of systems engineering and systems analysis that are both based on a systematic rational search for the best means/system of tackling problems and objectives, which are assumed to be clearly and objectively defined as technical specifications (Biazzo, 1998). They do not take into account other elements of the activities such as resources and rules. They also did not fit with the sociotechnical concept of systems but mainly examined the technical aspects of the process with a few mentioning but not taking into account social factors, or vice versa. Table 1 provides a summary of some of the reviewed literature.

Gap in Literature

After reviewing the literature, it became apparent that despite many attempts to map the processes of the OTWL, few took the requirements of the sociotechnical system theory into consideration. This includes

Table 1. Summary of some of the reviewed literature

Author	Title	Aim	Findings
Al-Hakim, 2006	Web-based hospital information system for managing operating theatre waiting list	To present a Web-based system for managing OTWL; there is the use of fuzzy logics	It is essential to map the process in order to identify the interdependencies between the process's elements.
McAleer et al., 1995	Simulation of a hospital's theatre suite	This study was undertaken to identify delays and restrictions in the system by using the VIS model to explore the relationship among various throughput levels and resource allocations.	There is no simple answer to the original problem of increasing throughput.
Buchanan & Wilson, 1996	Re-engineering operating theatres: the perspective assessed	Process re-engineering of the hospital work system based on analysis of the activity chain across the organization	Re-engineering perspective does not deliver clear solutions but provides a platform for further investigation.
Buchanan, 1998	Representing process: the contribution of a re-engineering frame	BPR to address the problems of scheduling patients to their OT and surgical teams to avoid both delays and overruns to the schedule.	BPR has many critics and weaknesses, but it can assist in being a tool to shape the definition of organization problems.
Foote et al., 2004	Toward a systematic understanding of a hospital waiting list	To find out why increased funding to clear waiting lists proves ineffective	Due to the social construction of the waiting list problem, translating research findings is unlikely to be straightforward.
NSW Health, 2002	Operating theatre management project	Define a set of appropriate performance indicators to flag variations in management and practice of operating theatres and to identify characteristics of operating theatre management that maximize effective theatre utilization	A conceptual framework of operating theatre management has been done.
Audit Commission, 2003	Operating theatres: a bulletin for healthy bodies	Explores common problems of operating theatres	There is an impetus for change, and good quality information is essential; an awareness of what is possible will help improve the situation.
Fulscher & Powell, 1999	Anatomy of a process mapping workshop	To share their experience in using IDEF0 in the workshop setting of redesigning the core auto insurance business	IDEF0 helped to focus the work of the group.
Biazzo, 2002	Approaches to business process: analysis	Highlights and compares alternative techniques and approaches of business process analysis	Process mapping plays an important role in helping to understand the structural dimensions of workflows. One of the reasons business process change is not successful lies in the fact that process designs pay insufficient attention to the social context of work.

the mechanism and control elements of the activities and the interdependencies as well as information flow between them. Thus, this is the gap in the literature that this research is attempting to fill. This chapter raises the following question: How can the flow of information and the interdependencies between the elements of the activities of the OTWL process be identified and mapped for the purpose of process re-engineering?

THEORETICAL FRAMEWORK

There are two ways to improve OTWL process: (1) through incremental changes, and (2) through breakthroughs (Evans & Lindsay, 2005). Incremental changes are small and improvements gradual in specific sections in OTWL. Breakthroughs are large or rapid improvements. Breakthroughs require redesigning the OTWL process; this is a reference to business process re-engineering (BPR) (Al-Mashari & Zairi, 1999). So far, most of the studies deal with the gradual change and improvement to the system (Audit Commission, 2002; Lismore, McAleer & Naqvi, 1995; NSW Health, 2002). There are limitations in the studies that deal with the re-engineering of emergency and OTWL process (Levary, 1997). This research looks to frameworks leading to re-engineering the OTWL process. The required initial step is to develop a framework that allows practicians to successfully implement the BPR.

As part of a sociotechnical system, each activity of the OTWL process has elements other then input and output. These additional elements are the resources that include information necessary to perform the activities and the rules that govern the activity's implementation. Therefore, the OTWL process is made up of complex sets of elements that are interdependent with each other. Mapping an OTWL process by breaking down the process into activities and tracing the input-output relationships of the activities is an essential step to improve the process, but it is not sufficient. An input-output relationship is only one type of interdependency between two types of elements. An activity within a sociotechnical system comprises elements other than input and output, and accordingly, there are interdependencies other than input-output relationships. An effective process mapping that allows us to improve or re-engineer the process also necessitates effective information mapping. This requires the following:

- Identifying the elements of each activity
- Determining the interdependencies among the elements of various activities

By determining the interferences, the information flow between the interdependent elements can be mapped. Coordination theory (Crowston, 1997; Lewis & Talalayevsky, 2004) offers a framework for understanding and managing the interdependencies. This research also finds that a structured process mapping technique referred to as IDEF0 (Fulsher & Powell, 1999; NIST, 1993) can be used effectively to identify elements of OTWL activities as well as the interdependencies among them. The identification of interdependencies among the elements of activities of the OTWL process allows us to coordinate the elements of the interdependencies and redesign the whole OTWL process.

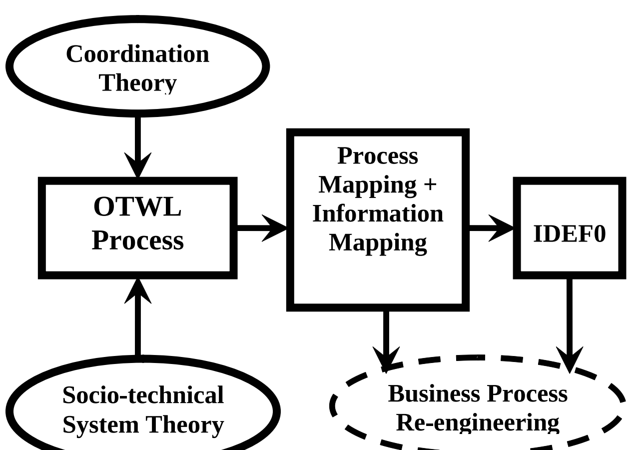
Figure 1 illustrates the developed theoretical framework. Following are brief descriptions of BPR and the related theories: sociotechnical systems theory and coordination theory.

Business Process Re-Engineering (BPR)

Business process re-engineering can be defined as the fundamental rethinking and radical redesign of business processes in order to achieve dramatic improvements in critical, contemporary measures of performance such as cost, quality, service, and speed (Al-Mashari & Zairi, 1999; Horna, 1995). The term "business process redesign" was first coined during a research program that started in 1984 at MIT (Biazzo, 2002).

In hospitals, it has been noted that during their stay, patients normally will come into contact with a range

Figure 1. Proposed theoretical framework



of professional staff in the hospital, each with their own management structures and hierarchies. This can result in a fragmented approach to care delivery, which fails to put the patient at the center of the process, thus disempowering the patient and providing less than effective care delivery (Buchanan & Wilson, 1996). Process re-engineering brings a holistic perspective to the hospital work system based on the analysis of relevant activities across the hospital, ignoring the functional boundaries and adopting a fresh start or a "blank-sheet" approach to process redesign. The cross-functional analysis of workflows in an organization and the novel ways of representing activity chains visually can be used to legitimatise and trigger a creative dialogue concerning ways to redesign work processes (Buchanan, 1998).

BPR is being considered for OTWL process problem because despite extensive studies being made in this area to understand the OTWL process, none of them has managed to adequately solve the problem. Only a process-based radical design of a sociotechnical system like the OTWL process is likely to succeed in improving the situation.

Sociotechnical Systems Theory

Sociotechnical Systems (STS) theory is a systems-based approach for process analysis and redesign (Fernandez et al., 2001). The term "sociotechnical" refers to the essential interdependence between the technical and the social systems. It was first coined in the 1960s by researchers in the Tavistock Institute of Human Relations, growing out of consulting work carried out while working on technical changes in coal mining, which was based on conceptual models from psychoanalysis (The Bayswater Institute, 2004; Sorensen, 1985). A systems-based approach provides a holistic consideration of the "problem system." This includes the social and technical elements, their formal and informal relationships, emergent patterns, and the unique context of the problem (Keating, Kauffman & Dryer, 2001).

Loughman, Fleck, and Snipes (2000) gathered the following insights that are pertinent to systems analysis from studying people in sociotechnical systems:

- People do not always use rational methods to make decisions; political methods can be as pervasive, if not more so.
- They view their world metaphorically and symbolically as well as literally.
- They will not always do what they are told to do.
- They can be very creative in sabotaging structures and processes they fear or dislike.
- The formal organization exhibited in the organizational chart inadequately describes how organization members actually conduct their business.

Therefore, it can be said that the internal system of any organization type consists of two distinct subsystems: the technical subsystem and the social subsystem, which relate, interdepend, and interact among themselves in a mutually inclusive fashion (Angelis & Antivachis, 1999).

The process mapping of the OTWL system should show the relationships among the activities, people, data, and objects (patients) involved in the production of a specified output. However, as already mentioned, the previous process mapping studies mainly examined the technical aspects of the situation with a few mentioning but not taking into account social factors. Under the social aspect of the OTWL process, the more important issues that are highlighted as found upon doing an analysis of the case studies are the following:

- **The behavior of the object (patient) is not predictable, and it varies.** This has been found from the case study with a hospital. For example, a patient's cancellation of surgery at the last minute is very common, resulting in much waste of resources and the need to reschedule the patient, thus causing further delay.
- **Surgeon effectiveness variability.** It is hard to measure how effective surgeons perform each and every surgery, because every surgeon differs in terms of skills and expertise. Effectiveness is using a skill in a way that raises the probability of success. Effectiveness here could be measured in terms of success.
- **Variability of the degree of success of the operation.** The success of the operation is hard to determine before the operation because of reasons such as variations in a patient's condition, which cannot be fully assessed until a surgical procedure is underway.
- **Variability of surgical time.** The resultant unpredictability of surgical time renders difficult any attempt at precision scheduling of theatre lists.

Therefore, it can be said that OTWL consists of not just the technical aspect but also the social aspects, and the STS theory is used to address this issue.

Coordination Theory

Coordination theory provides part of the theoretical background to this research. Its intent is to analyze organizations in a way that facilitates redesign (Crowston, 1997). Coordination theory offers a framework for understanding and characterizing various types of dependencies and identifies the processes that can be used to manage these dependencies (Lewis & Talalayevsky, 2004). The aim of this theory, therefore, is to define the interdependencies of the processes and attempt to improve performance. Coordination problems arise from dependencies that constrain the performance of the tasks (Crowston, 1997). Given this perspective, alternative processes would be identified for performing a desired task.

In this research, the OTWL management process can be described as a kind of producer/consumer relationship that occurs when the output of one step in the process sequence is the input to the next. Producer/consumer relationships can lead to the following kinds of dependencies (Malone & Crowston, 1994):

- **Prerequisite constraints.** A producer activity and a consumer activity are dependent because the producer activity must be completed before the consumer activity can begin. This is similar to the Coordination Theory idea of a sequential interdependency.
- **Transfer:** When one activity produces something that is used by another activity, the thing produced must be transferred from the producer activity to the consumer activity. The transfer, be it physical or just communication of information, is a coordination activity. This can be defined as a reciprocal interdependency under the Coordination Theory.

Table 2. A typology of dependencies and associated coordination mechanisms (Crowston, 1997)

Interdependence	Coordination Activity
Task-task	
Tasks share common output	
Same characteristics	- look for duplicate tasks - merge tasks or pick one to do
Overlapping	negotiate a mutually agreeable result
Conflicting	pick one task to do
Tasks share common input (shared resource)	
Shareable resource	no conflict
Reusable resource	- notice conflict - schedule use of the resource
Nonreusable resource	pick one task to do
Output of one task is input of other (prerequisite)	
Same characteristics	- order tasks - ensure usability of output - manage transfer or resources
Conflicting	- reorder tasks to avoid conflict - add another task to repair conflict
Task-resource	
Resource required by task	- identify necessary resources - identify available resources - choose a particular set of resources - assign the resources
Resource-resource	
One resource depends on another	- identify the dependency - manage the dependency

The OTWL management process comprises all the interdependencies mentioned so far, including the pooled, sequential (prerequisite constraint), reciprocal (transfer), and task/subtask dependencies.

Dependencies can be between tasks, between tasks and the resources they need, or between the resources used (Crowston & Kammerer, 1998). Table 2 shows examples of generic coordination activities in relation to the type of interdependencies. The classification is based on tasks and resources. Table 2 implies the importance of identifying the interdependencies in order to determine the information flow within and among these interdependencies.

USING IDEF0 TO BUILD PROCESS/INFORMATION MAPPING

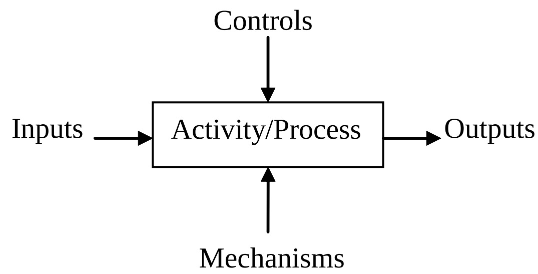
IDEF0 is one part of a comprehensive process definition and computer systems implementation methodology that was originally developed for the United States Air Force (Fulsher & Powell, 1999). It is widely used in process mapping and is, in fact, the official standard of the U.S. government for this purpose (Bal, 1998). It is a modeling technique based on combined graphics and text that are presented in an organized and systematic way to gain understanding, support analysis, and provide logic for potential changes. It is a highly structured and disciplined framework within which logical and consistent process maps can be drawn to any desired level of detail (NIST, 1993). Figure 2 shows generically how IDEF0 is used to depict activities and their associated elements: inputs, outputs, controls, and mechanisms.

In IDEF0, the components of the IDEF0 syntax are boxes and arrows, rules and diagrams. The boxes represent functions, also known as activities or processes. Arrows represent data or objects related to the activity.

The Elements of an Activity

The result of applying IDEF0 is a model that consists of diagrams, text, and glossary, all cross-referenced to each other. The box and arrow meanings are used to relate several subfunctions on a diagram comprising a more general function. Inputs are data or objects that are consumed or transformed by an activity; they are represented by arrows entering the left side of the box.

Figure 2. IDEF0 elements



Outputs are data or objects that are the direct results of an activity; they are shown as arrows leaving the right side of the box. In this research, the OTWL processes are described as a set of linked activities, each with inputs and outputs. External or internal factors control each activity. Controls are data or objects that specify conditions that must exist for an activity to produce correct outputs and is represented by arrows entering the top of the box. It can also be said that the controls of an activity refer to the actions, policies, or regulations that direct the activity. Each activity requires one or more mechanisms and/or resources. They support the successful completion of an activity but are not changed in any way by it. Mechanisms are represented in the graphical diagram by arrows connected to the bottom side of the box. It is assumed that each activity may have several inputs and outputs, and may be embraced by a number of controls and may require several mechanisms (Al-Hakim, 2006; Fulscher & Powell, 1999; NIST, 1993). IDEF0 allow the user to "tell the story" of what is happening in the system (Liyanage & Perera, 2001).

Creating Process Maps with IDEF0

IDEF0 is chosen over competing methods because it uses a hierarchical top-down approach and focuses the discussion on the conceptual basis for a process rather than on the detailed sequence of specific microlevel activities required to deliver the end result (Fulscher & Powell, 1999); therefore, it suits the multicriteria objective of the operation theatre waiting list management process. The hierarchical series of diagrams gradually display increasing levels of detail describing functions and their interfaces within the context of a system. There are three types of diagrams: graphic, text, and glossary. The graphics diagrams define functions and functional relationships via box and arrow syntax and semantics as seen in Figure 3. The figure also shows that the primary hierarchical relationship is between a parent box and the child diagram that details it.

A box in an IDEF0 model represents the boundaries drawn around some activity. Looking inside the parent box leads to the discovery of the breakdown of the activity into smaller activities as seen in the child diagram, which together comprises the box at the higher level. This is useful for hiding unnecessary complexity from view until a more in-depth understanding is required by looking inside the box at its decomposition (Mayer, Painter, & deWitte, 1992).

IDEF0 has been most commonly used in the visualization of manufacturing systems (Bal, 1998). There have been a few attempts to use it in the hospital context (Chick, Sanchez, Ferrin & Morrice, 2003) but not in the area of the surgical waiting lists.

Decomposition is generally necessary for any improvement or intervention as only a thorough decomposition allows for a pervasive recomposition like the re-engineering of the OTWL process (Van Der Zwaan & De Vries, 2000). IDEF0 allows such decomposition of a system's activities to take place. Also, the interactions among activities are not only in the form of inputs and outputs, but in many cases the output of an activity may control another activity or may form a feedback that requires the reworking of a prior activity. As seen in Figure 4, the output of Function A forms input into both Function B and C, while the output of Function B controls Function C, while at the same time the output of Function C controls Function B (NIST, 1993). This is a constructive example of information mapping in which information resulted from an activity (Activity C) can control another one (Activity B). Figure 4 shows that information control activity A is also control Activity B.

Identifying Interdependencies with the IDEF0 A/C Matrix

Once the processes are mapped in IDEF0, the software also produces an Activity/Concept (A/C) matrix that effectively identifies the interdependencies among the inputs, control, outputs, and mechanisms, both within

Figure 3. Example of hierarchical top-down model (NIST, 1993)

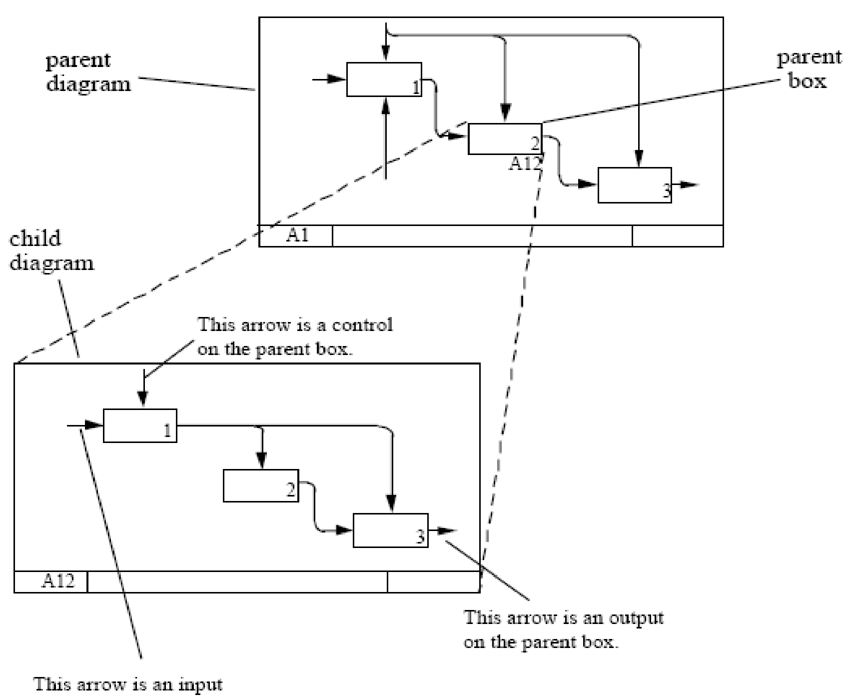
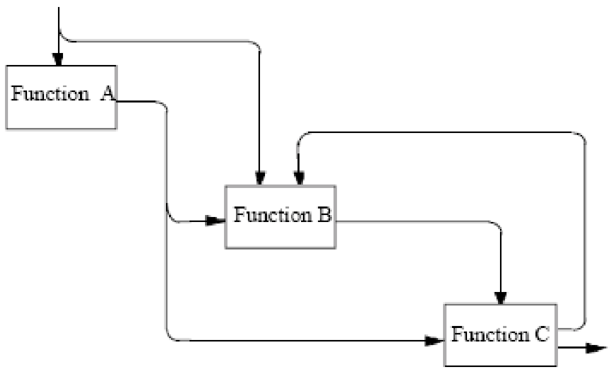


Figure 4. Interdependencies between activities (NIST, 1993)



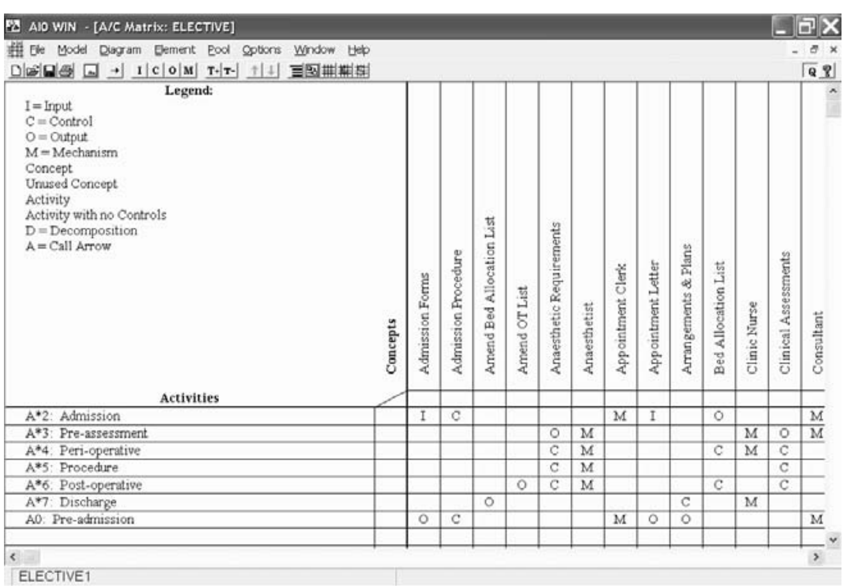
and between the activities (see Figure 5 for an example A/C matrix). This matrix shows the relationships among the activities' elements in terms of input (I), output (O), control (C), and mechanism (M). These relationships form the other facet of information mapping between these elements.

Classifying Interdependencies from IDEF0 Process Maps

Al-Hakim (2006) identified the following six basic forms of interdependencies among the elements of activities or tasks represented in an IDEF0 process map. Each numbered dependency type corresponds to the same number as seen in the diagram in Figure 6.

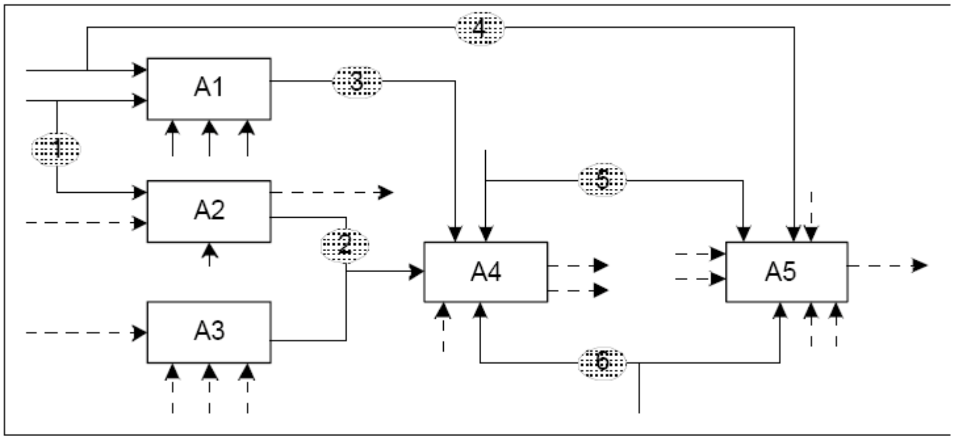
1. **Input/input dependency.** The same input is required for multiple activities. This can be seen in how the arrow representing the use of the input branches out into two arrows into the activities A1 and A1 respectively. They use the same input and thus it can be said that there is an input/input interdependency between them.
2. **Output/input dependency.** The output from one activity acts as input to another activity and can be seen in how the arrow representing the output of A2 becomes an input for activity A4.
3. **Output/control dependency.** The output from one activity acts as a control for another activity. In Figure 6, the output of activity A1 forms a control for A4; thus, there is output/control interdependency between activity A1 and activity A4.
4. **Input/control dependency.** The input for one activity acts as a control for another activity. This can be seen in how an input for activity A1 is a control for activity A5 as well, and so there is input/control dependency between these two activities.
5. **Control/control dependency.** The same control applies to multiple activities as seen in how both activities A4 and A5 are under the same con-

Figure 5. Activity/concept matrix



Activities / Concepts	Admission Forms	Admission Procedure	Amend Bed Allocation List	Amend OT List	Anaesthetic Requirements	Anaesthetist	Appointment Clerk	Appointment Letter	Arrangements & Plans	Bed Allocation List	Clinic Nurse	Clinical Assessments	Consultant
A*2: Admission	I	C					M	I		O			M
A*3: Pre-assessment					O	M					M	O	M
A*4: Peri-operative					C	M				C	M	C	
A*5: Procedure					C	M						C	
A*6: Post-operative				O	C	M				C		C	
A*7: Discharge			O						C		M		
A0: Pre-admission	O	C					M	O	O				M

Figure 6. Six forms of interdependencies representing the flow of information mapping (Al-Hakim, 2006)



trol.

6. **Resource/resource dependency.** The same resource is used by multiple activities. Activities A4 and A5 both utilize the same resources, as seen in the following figure, and so they have a resource/resource dependency.

The mentioned forms of interdependencies in this section are those that can be identified from the result of IDEF0 mapping.

CLASSIFICATION OF INTERDEPENDENCIES

A comparison of the interdependencies identified by Crowston (1998) in Table 2 and by Al-Hakim (2006) in Figure 6 reveals that while the two lists overlap to some degree, there are also differences in both content and the use of terms. For the purpose of classifying interdependency types, the two lists are compared and contrasted, with a summary presented in Table 3.

All six of Al-Hakim's (2006) interdependency types are classified from the activity/task viewpoint, and therefore, all six are task-task interdependencies. Also, Crowston's (1998) typology of interdependencies does not include the influence of controls at all, nor does Crowston differentiate between inputs (data or objects consumed or transformed by the activity) and resources (mechanisms or objects that support the completion of the activity but that are not consumed by the activity). The resource-resource interdependency identified by Crowston is different in context and meaning to the resource-resource interdependency identified by Al-Hakim. Crowston uses the term resource-resource dependency to refer to the situation when one resource is dependent on another resource within the same activity. While this situation may occur within the OTWL process, the A/C matrix does not include any way of showing this type of dependency at all. The best the matrix can do is to show the task-resource dependencies; that is, identify all resources needed for the same task. Al-Hakim uses the term resource-resource dependency to mean that at least two or more tasks are dependent on the same resource (see number 6 in Figure 6), and this meaning will be adopted in this research.

Thus, there are two major groups of interdependencies: (1) dependencies among activities; and (2) dependencies within an activity. Table 3 summarizes the interdependency types identified by both Crowston (1997) and Al-Hakim (2006). Table 3 also illustrates that all but one of the interdependency types can be located using the IDEF0 A/C matrix by either reading down a column, or reading across a row.

For example, Figure 3.6 shows output/control dependencies in the Clinical Assessments column, the Anaesthetic Requirements column, and the Bed Allocation List column. A task-resource interdependency occurs in the pre-assessment activity (A3) with the anaesthetists, clinical nurses, and consultants all being required resources for this activity to function. A resource-resource dependency occurs with the anaesthetist being required for four activities: pre-assessment, peri-operative, procedure, and post-operative.

Table 3. Comparison of Crowston's and Al-Hakim's type of interdependencies

	Task–Task (interdependency *among* tasks)	**Crowston (1997)**	**Al-Hakim (2006)**	**A/C Matrix Location**
1		Tasks share common output	-	Column
2		Output of one task is input of other	Output/Input (or vice versa) dependency	Column
3		Tasks share common input	Input/Input dependency	Column
4		Tasks share common input	Resource/Resource dependency	Column
5		-	Output/Control dependency	Column
6		-	Input/Control dependency	Column
7		-	Control/control dependency	Column
8	***Task–Resource*** (interdependency *within* a task)	Resource required by task	-	Row
9	***Resource–Resource*** (interdependency *within* a task)	One resource depends on another	-	Not represented in A/C matrix

Limitation of IDEF0

Mapping with IDEF0 comprises some limitations. IDEF0 does not include representation of time. Also, the logical flow of the elements and activities is not really meant to depict the sequence of events exactly. The hierarchical decomposition of activities means that we can only analyze each level and section on its own; the software does not allow cross-section/-level analysis in itself.

CONCLUSION

The OTWL process is a very complex and dynamic process within the hospital, which itself is a complex sociotechnical system. Problems related to growing operating theatre waiting lists are getting worse, and there are many reasons behind this. Despite much research and many attempts to solve the problem using flowcharts and other forms of process mapping, no previous study has taken the requirements of the sociotechnical system theory into consideration by including the resource/mechanism and control elements of the activities and by identifying the interdependencies that exist among all four input, control, mechanism, and output elements of the OTWL process. Identifying the interdependencies provides a picture of information flow within and among the elements of the activities. Using IDEF0, and with the concept of BPR and coordination theory as well, this research tries to address this problem.

The theoretical framework of this research is deemed appropriate for the topics of mapping the sociotechnical system of the OTWL process because of the following issues (Foo & Al-Hakim, 2006):

1. It takes into account the social and technical aspects of a process
2. It identifies and maps the four elements (input, output, resource, control),
3. As well as the interdependencies of the activities/elements and the flow of information between the elements

4. It is easy to identify the flow of information and interdependencies from the graphic description of the systems in IDEF0
5. There is also the allowance for an activity to form a feedback that requires the reworking of a prior activity
6. The coordination theory allows us to see if the elements are linked effectively, giving us the opportunity to change the relationship of the elements and to create new interdependencies; thus, the process can be redesigned.

As a result, this theoretical framework is useful for redesigning a process. Therefore, the research question—How can the flow of information and the interdependencies among the elements of the activities of the OTWL process be identified and mapped for the purpose of process reengineering?—has been effectively addressed.

ACKNOWLEDGMENT

The research was partially sponsored by the University of Southern Queensland. The author would like to acknowledge the thoughtful time of Queensland Health Project Liaison Officers, Ms. Lee Hunter, Mr. Brett Mendezona, and Ms. Sylvia Johnson.

REFERENCES

Al-Hakim, L. (2006). Web-based hospital information system for managing operating theatre waiting list. *International Journal of Healthcare Technology and Management*, *7*(3/4), 266–282.

Al-Hakim, L., & Fitzgerald, A. (2003). Web-mobile-based technologies: A fuzzy logics approach to decision-making about queues in operating theatre suites. In *Proceedings of the 4th International Web Conference*, Perth, Australia.

Al-Mashari, M., & Zairi, M. (1999). BPR implementation process: An analysis of key success and failure factors. *Business Process Management Journal*, *5*(1), 87–112.

Angelis, V.A., & Antivachis, N.A. (1999). The organizational performance re-considered: The case of the technical and social subsystems synergy. *Business Process Management, 5*(3), 275–288.

Audit Commission. (2002). *Operating theatres: A bulletin for health bodies*. Retrieved October 3, 2005, from http://www.audit-commission.gov.uk/national-studies/publishedstudies-health.asp

Audit Commission. (2003). Operating theatres: Review of national findings. Retrieved July 3, 2005, from http://www.audit-commission.gov.uk/reports/AC-REPORT.asp?CatID=PRESS-CENTRE&fromPRESS=AC-REPORT&ProdID=6CDDBB00-9FEF-11d7-B304-0060085F8572

Bal, J. (1998). Techniques: Process analysis tools for process improvement. *The TQM Magazine*, *10*(5), 342–254.

Barlow, G. (2002). Auditing hospital queuing. *Managerial Auditing Journal*, *17*(7), 397–403.

The Bayswater Institute. (2004). Sociotechnical Systems Design Workshop. Retrieved October 11, 2005, from http://www.bayswaterinst.org/subpage1.html

Biazzo, S. (2002). Process mapping techniques and organisational analysis: Lessons from sociotechnical system theory. *Business Process Management*, *8*(1), 42–52.

Buchanan, D. (1998). Representing process: The contribution of an engineering framework. *International Journal of Operations and Production Management*, *18*(12), 1163–1188.

Buchanan, D., & Wilson, B. (1996). Re-engineering operating theatres: The perspective assessed. *Journal of Management in Medicine*, *10*(4), 57–74.

Buhaug, H. (2002). *Long waiting lists in hospitals: Operational research needs to be used more often and may provide answers*. Retrieved October 20, 2005, from http://bmj.bmjjournals.com/cgi/content/full/324/7332/252

Carvel, J. (2001). Patient care hit by waiting list policy. *Guardian Newspapers*. Retrieved July 20, 2005, from http://society.guardian.co.uk/nhsperformance/story/0,8150,527664,00.html

Chick, S., Sanchez, P.J., Ferrin, D., & Morrice, D.J. (2003). Redesigning the medication ordering, dispensing and administration process in an acute care academic health sciences centre. In *Proceedings of the 2003 Winter Simulation Conference*.

Chua, K.P. (2005). Waiting lists in Canada: Reality or hype? Retrieved October 20, 2005, from http://www.amsa.org/studytours/WaitingTimes_primer.pdf

Council of Europe. (2005). *Social cohesion and quality of life: Criteria for the management of waiting lists and waiting times in health care.* Retrieved July 20, 2005, from http://www.coe.int/T/E/Social_Cohesion/Health/Documentation/Waiting%20lists%20report%20&%20%20rec%20PUB%20English.asp

Crowston, K. (1997). A coordination theory approach to organizational process design. *Organization Science, 8*(2), 157–175.

Crowston, K., & Kammerer, E.E. (1998). Coordination and collective mind in software requirements. *IBM Systems Journal, 37*(2), 227–246.

Evans, J.R., & Lindsay, W.M. (2005). *The management and control of quality* (6th ed.). Minneapolis/St. Paul: West Publishing.

Fernandez, A.A., Jacobs, D.A., Kauffman, P., & Keating, C.B. (2001). A methodology for analysis of complex sociotechnical processes. *Business Process Management Journal, 7*, 33–49.

Foo, G., & Al-Hakim, L. (2006). Theoretical framework for mapping information of social technical process. In *Proceedings of the 2006 Information Resources Management Association International Conference*, Washington, D.C., 719–723. Washington DC.

Foote, J.L., North, N.H., & Houston, D.J. (2004). Towards a systemic understanding of a hospital waiting list. *Journal of Health Organization and Management*, *18*(3), 140–154.

Fulscher, J., & Powell, S.G. (1999). Anatomy of a process mapping workshop. *Business Process Management Journal, 5*(3), 208–237.

Gonzalez-Busto, B., & Garcia, R. (1999). Waiting lists in Spanish public hospitals: A system dynamics approach. *System Dynamics Review, 15*(3), 201–224.

Healthcare Commission. (2005). *Acute hospital portfolio review: Day surgery*. Retrieved October 20, 2005, from http://www.healthcarecommission.org.uk/NationalFindings/OtherPublications/fs/en?CONTENT_ID=4017471&chk=XdgkyO

Horna, P. (1995). Business process re-engineering: Theory and evidence-based practice. *Business Process Reengineering & Management Journal, 1*(3), 10–30.

Ivaldi, E., Elena, E.T., & Testi, A. (2003). *Simulation supporting the management of surgical waiting lists.* Retrieved October 10, 2005, from http://www.economia.unige.it/04/diem/new/economia/wp/21.pdf

Johnson, J.A, Jones, W.J., & Siddharthan, K. (1996). A priority queuing model to reduce waiting times in emergency care. *International Journal of Health Care Quality Assurance, 9*(5), 10–16.

Keating, C.B., Kauffman, P., & Dryer, D. (2001). A framework for systemic analysis of complex issues. *Journal of Management Development, 20*(9), 772–784.

Levary, R.R. (1997). Re-engineering hospital emergency rooms: An information system approach. *International Journal of Health Care Quality Assurance*, *10*(5), 179–191.

Lewis, I., & Talalayevsky, A. (2004). Improving the interorganizational supply chain through optimization of information flows. *The Journal of Enterprise Information Management, 17*(3), 229–237.

Lismore, T.D., McAleer, W.E., & Naqvi, I.A. (1995). Simulation of a hospital's theatre suite. *Journal of Management in Medicine, 9*(5), 14–26.

Liyanage, K., & Perera, T. (2001). IDEF based methodology for rapid data collection. *Integrated Manufacturing Systems, 12*(3), 187–194.

Loughman, T.P., Fleck, R.A.J., & Snipes, R. (2000). A cross-disciplinary model for information systems analysis. *Industrial Management & Data Systems*, *100*(8), 359–369.

Malone, T.W., & Crowston, K. (1994). The interdisciplinary study of coordination. *Computing Surveys, 26*(1), 87–119.

Mayer, R.J., Painter, M.K., & deWitte, P.S. (1992). *IDEF family of methods for concurrent engineering and business re-engineering applications*. TX: Knowledge Based Systems, Inc.

McAleer, W.E., Turner, J.A., Lismore, D., & Naqvi, I.A. (1995). Simulation of a hospital theatre suite. *Journal of Management in Medicine*, *9*(3), 14–26.

National Institute of Standards and Technology (NIST). (1993). Federal information processing standards publication 183, specifications for integration definition for function modelling (IDEF0). Retrieved May 15, 2005, from http://www.nist.gov/itl/lab/fips/index.html

NSW Health. (2002). *Operating theatre management project—Supplementary report on process mapping*. NSW Health Department.

Ramirez-Valdevia, M.T., & Crowe, T.J. (1997). Achieving hospital operating objectives in the light of patients. *International Journal of Health Care Quality, 10*(5), 208–212.

Schofield, W.N., et al. (2005). Cancellation of operations on the day of intended surgery at a major Australian referral hospital. *Medical Journal of Australia*, *182*(12), 612–615.

Scottish Executive. (2001). Waiting times to become litmus test of NHS performance. August 30, 2001.

Soliman, F. (1998). Optimum level of process mapping and least cost business process re-engineering. *International Journal of Operations and Production Management, 18*(9-10), 810–816.

Sommerfeld, J. (2005). Waiting lists blow out as beds cut. *The Courier-Mail*, September 21, 2005, p. 2.

Sorensen, K.H. (1985). Technology and industrial democracy: An inquiry into some theoretical issues and their social basis. *Organization Studies*, *6*(2), 139–160.

Van Der Zwaan, A.H., & De Vries, J. (2000). A critical assessment of the modern sociotechnical approach within production and operations management. *International Journal of Production Research, 38*(8), 1755–1767.

Master-Slave Robotic System for Therapeutic Gastrointestinal Endoscopic Procedures

S.C. Low
Nanyang Technological University, Singapore

S.J. Phee
Nanyang Technological University, Singapore

S.W. Tang
Nanyang Technological University, Singapore

Z.M. Thant
Nanyang Technological University, Singapore

K.Y. Ho
National University of Singapore, Singapore

S.C Chung
National University of Singapore, Singapore

INTRODUCTION

Flexible endoscopy is used to inspect and treat disorders of the gastrointestinal (GI) tract without the need for creating an artificial opening on the patient's body (Phee, Ng, Chen, Seow, & Davies, 1997). The endoscope is introduced via the mouth or anus into the upper or lower GI tracts respectively. A miniature camera at the distal end captures images of the GI wall that help the clinician in diagnosis of the GI diseases. Simple surgical procedures (like polypectomy and biopsy) can be performed by introducing a flexible tool via a working channel to reach the site of interest at the distal end. The types of procedures that can be performed in this manner are limited by the lack of maneuverability of the tool. More technically demanding surgical procedures like hemostasis for arterial bleeding, suturing to mend a perforation, and fundoplication for gastrooesophageal reflux cannot be effectively achieved with flexible endoscopy. These procedures are often presently being performed under opened or laparoscopic surgeries.

With the invention of medical robots like the Da Vinci (Intuitive Surgical Incorporation) surgical systems, clinicians are now able to maneuver surgical tools accurately and easily within the human body (Carrozza, Dario, & Phee, 2003). Operating from a master console, the clinician is able to control the movements of laparoscopic surgical tools real time. These tools (also known as the slaves) are designed with sufficient degrees of freedom to move according to the natural hand and wrist motion allowing the clinician to perform intricate procedures with minimal technical difficulties.

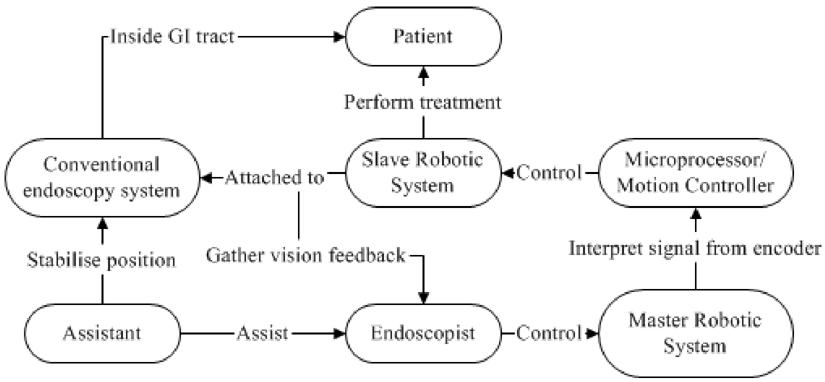
Thus far, master-slave surgical robotic systems like Zeus and Da Vinci are either completely rigid or do not have a significant length of flexible body (Low & Phee, 2004). The slave manipulators enter the human body by means of incisions on the body. In this article, we propose a robotic system with flexible slave manipulators that could be attached directly to flexible endoscopes. Similar to a conventional flexible endoscopic procedure (e.g., gastroscopy), the "robotic endoscope" bundle could negotiate the curves and bends in the GI to reach the desired position within the gut. This system empowers the surgeons to perform more difficult surgeries that are otherwise impossible with the conventional endoscopic tool.

OVERALL ROBOTIC SYSTEM

Figure 1 shows the proposed system layout whereby the endoscopist work on the master console while gathering visual feedback from the endoscope. A

M

Figure 1. Proposed system layout for the overall robotic system



computer console will interpret the readings from the master console that will in turn give instruction to the slave robotic system to perform the treatment to the patient. This system allows complicated treatment to be performed with the added benefit of easy and intuitive control for the endoscopist.

THE SLAVE MANIPULATOR

The 3D model of the intended slave manipulator can be seen in Figure 2. In order for the surgeon to perform the necessary dexterous actions, the slave manipulators should possess a high number of Degrees of Freedom (DOF). The emphasis of the project is to make the slave manipulator to be as intuitive to control as possible. As such, the DOF and joints of the slave manipulator are modeled after a simplified human arm as shown in Figure 3. Altogether there are five DOF for positioning of the slave and an extra DOF for manipulating the end effector and the axis or rotation. Two slave manipulators are used instead of one since it can perform actions such as pulling and cutting of polyps or suturing bleeding sites. Furthermore two slave manipulators are more intuitive to use since they resemble the two human arms.

In order for the slave manipulator to be able to go through human GI tract, the slave manipulator has to be small yet flexible. Due to this unique requirement, tendon-sheath actuation is used. The prime movers are located outside the human body and it transmits power to the mechanism by pulling and releasing tendons in a sheath accordingly.

Figure 2. 3D model of slave manipulators attached with the endoscope

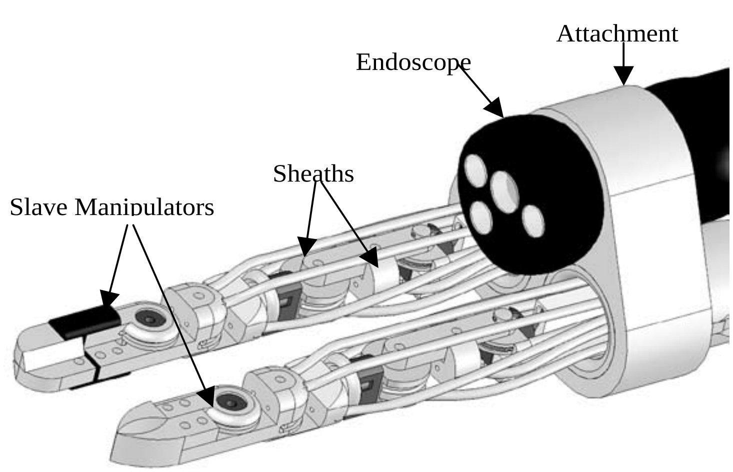
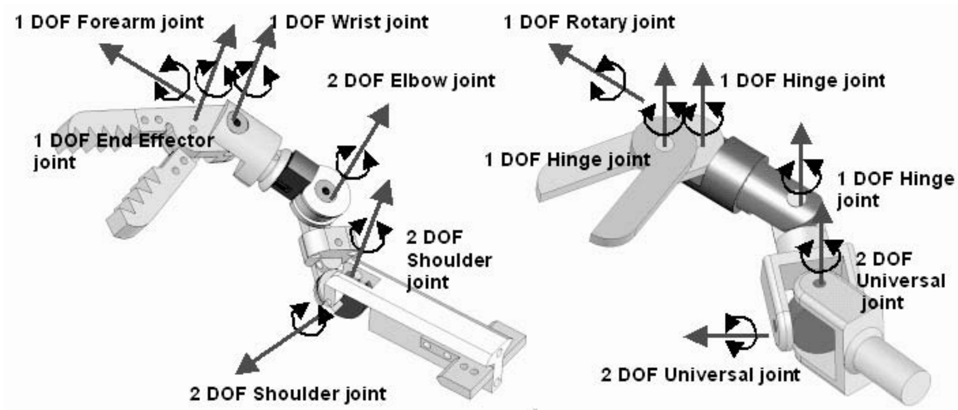


Figure 3. The DOFs for one arm of the slave manipulator together with its simplified joint diagram



The design of the mechanism is such that the two antagonistic tendons that control each DOF motion are independently actuated by one direct current (DC) motor. In addition, the rotational displacement of each joint is directly proportional to the linear displacements of the tendon. These characteristics allow the serial slave manipulator to be controlled more easily. With the help of the software, the manipulator will not face singularity conditions in the prescribed range of motion for the manipulators.

Different types of end effectors have been designed to perform different types of actions during surgery. One of the tools includes electrodes that can perform monopolar cauterizing. The picture of the current developed prototype can be seen in Figure 4. Take note that the attachment that held the endoscope and manipulators is not fixed to the current prototype yet.

To pass the endoscope and slave system to the desired site of the GI tract, an overtube can be used. Currently the size of the whole slave manipulator is approximately 25mm in diameter. Future developments will be on the miniaturizing of the slave manipulator till it is comfortable to be used on human beings.

THE MASTER CONTROLLER SYSTEM

The ergonomic master device is built as a 6 DOF structure to control the 6 DOF slave mechanism. As the slave manipulators resemble the human arms, the anthropomorphic data of the surgeon's arm is used as the control parameters in order to give the surgeon better perception in performing the joint-to-joint control of the slave.

The schematic of the master controller is shown in Figure 4. Two cable-actuated position sensors are used to measure the two contributing rotations of the shoulder joint that cause the up-down and in-out movement of the upper arm. Two optical rotary encoders are used to measure the flexion and supination-pronation movement of the elbow joint. Another similar encoder is used for flexion-hyperextension movement of the wrist, and the finger grip comes with an encoder inside the holder.

A virtual plane is formed using the reference of the shoulder joint with the two position sensors. The sensor locations can be modified to meet the required workspace that lies on one side of the plane. For the full range of upper arm movement, an additional sensor will be needed. As the comfortable workspace for

Figure 4. Figure of the current slave prototype for the robotic system

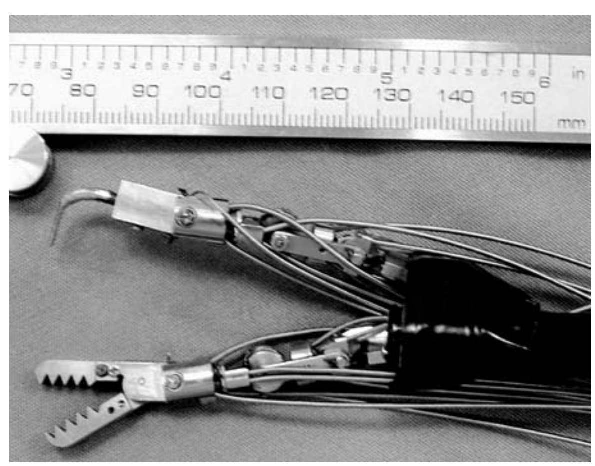
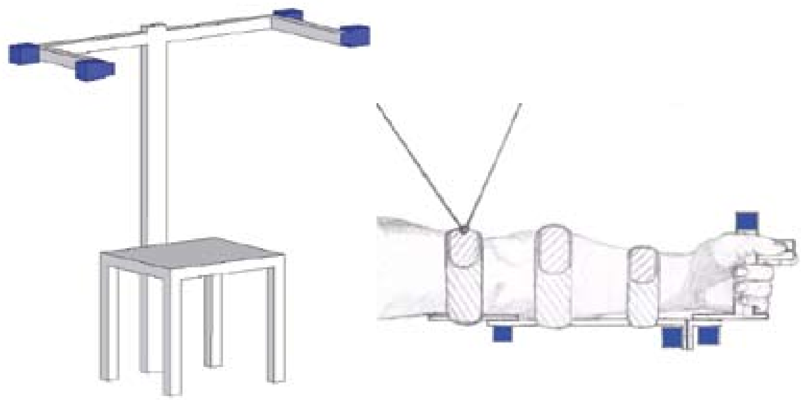


Figure 5. The whole schematic of the master setup (a) frame with two cable sensors for each arm (b) left arm with sensors attached



the surgeon's upper arm lies in a quarter of a spherical space, two position sensors are sufficient to get the anthropomorphic data of one arm in that range.

The position sensors are located in such a way that the cables do not impose any disturbance on the surgeon's vision and movement. Furthermore the tensions of the cables facilitate the surgeon as a counterweight from his own arm moments, reducing the fatigue.

Assuming the shoulder joint to be a ball and socket type, the point of cable-attached upper arm is determined by three spheres formed by upper arm length and two cable lengths. In Figure 6, L_1 and L_2 are the two cable lengths and L_3 is the upper arm length.

In $X_0Y_0Z_0$ coordinate system,

$$\left({}^0P_x - {}^0P_{1x}\right)^2 + \left({}^0P_y - {}^0P_{1y}\right)^2 + \left({}^0P_z\right)^2 = L_1^2 \qquad (1)$$

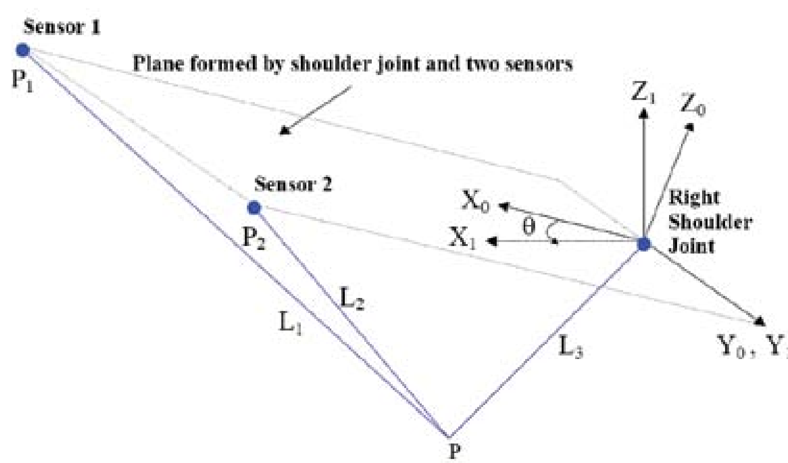
$$\left({}^0P_x - {}^0P_{2x}\right)^2 + \left({}^0P_y - {}^0P_{2y}\right)^2 + \left({}^0P_z\right)^2 = L_2^2 \qquad (2)$$

$$\left({}^0P_x\right)^2 + \left({}^0P_y\right)^2 + \left({}^0P_z\right)^2 = L_3^2 \qquad (3)$$

The point ${}^0P(x,y,z)$ below the plane can be solved from above equations and it can be transformed into $X_1Y_1Z_1$ by rotation matrix about Y-axis.

$$\begin{Bmatrix} {}^1P_x \\ {}^1P_y \\ {}^1P_z \end{Bmatrix} = \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix} \begin{Bmatrix} {}^0P_x \\ {}^0P_y \\ {}^0P_z \end{Bmatrix} \qquad (4)$$

Figure 6. The location of cable-attached upper arm in two coordinate systems



[1]P(x,y,z) can be used to calculate the shoulder joint up-down and in-out angles.

The mobility range of the master is constrained only at the upper arm by the virtual plane formed by the shoulder joint and two cable sensors. The other joints accommodate the full range of motion of the surgeon.

SOFTWARE CONTROL

A dedicated computer software is programmed to control and interface the slave and master robotic system. First it takes in real time readings from each encoder of the master device to determine the motion generated by the user. This data is processed and appropriate signals are sent out to actuate the respective motors to bring about the required movements of the slave. The system framework adopts a two-layer architecture. The bottom layer is the hardware control module that uses low-level drivers for hardware control. The top layer modules are the kernel algorithms, which perform the signal processing to compensate for backlash and noise disturbances.

PRELIMINARY EXPERIMENTS

In vitro experiments with the master-slave system have been conducted in carrying out some tasks. By handling the master device, the surgeons are able to control the slave manipulators to perform tasks such as pick and place as well as grab and cut a soft object. In the future, animal trials will be conducted to verify its functionality. During experimentation, it was found that the mechanism is capable of producing at least 1.0N for each DOF.

DISCUSSION

One of the biggest problems in the tendon-sheath driving robot system control is the nonlinear characteristic due to the friction between the tendon and sheath (Kaneko, Paetsch, & Tolle, 1992, pp. 96)

As a result, delays and movement hysteresis are noticeable and can be seen in the following relationship:

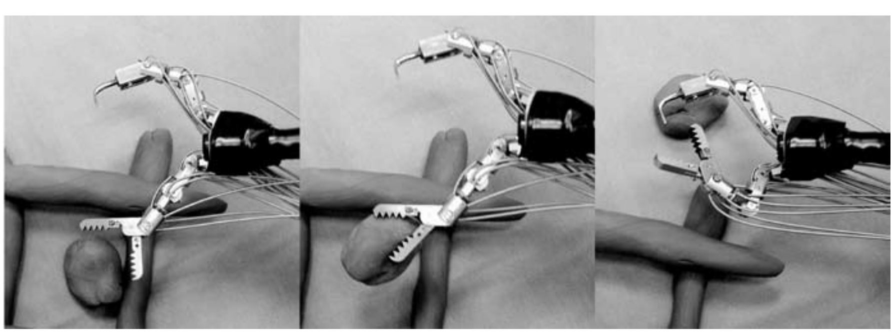
$$\delta(L) = \frac{T_0 L}{EA} \frac{\exp(\lambda) - \lambda - 1}{\lambda} \quad (5)$$

Where T_0 is tendon-pretension, L is the length of the tendon-sheath, E is Young's modulus, A is the cross section area of tendon, λ is a nondimensional parameter and indicates the total friction force acting on the tendon under unit tendon-tension. We can regard $\delta(L)$ as a kind of backlash for the tendon-sheath driving system. For any tendon-sheath driving system, L, E, and A can be considered as constant. λ can also be considered as a constant if the curve of the tendon-sheath does not vary much. From (5), the backlash length is only a function of the cable tension T_0. To improve the performance of the system, a pretension device is introduced so that the tendon is always in tension. The modified mechanism from (Hirose & Chu, 1999) can also minimize the problem of wire elongation after repeated use.

The pretension is set in such a way that the tension change during robot movement is trivial compare to the pretension, so that backlash length $\delta(L)$ can be considered as a constant.

To further reduce the movement delay of the slave, the software will record the movement of each slave joint and whenever a direction change is detected, additional actuator displacement is used to compensate

Figure 7. Using the robot to perform pick and place tasks



M

for the backlash. Experiments have been designed and performed to obtain the appropriate backlash compensation displacement for each joint. Using this method, the delay of the slave can be reduced by up to 0.5 second. The system can be improved in the future with the modeling of the system as well as making use of force sensors on the slave robot.

FUTURE TRENDS

To minimize the size of the slave manipulator while maintaining a high force, it is necessary to study and improve on the current tendon sheath system. This is because it has been found that the bulk of the force is lost as friction between the long and flexible tendon and sheath.

Another important aspect that needs to be implemented is haptics. To achieve haptics, there must be knowledge of the force experienced by the slave manipulator. However fixing sensors on the small slave manipulator is a difficult task and there are currently studies made to obtain the information of the force experienced by the slave manipulator.

CONCLUSION

A master slave robotic system that can enhance gastrointestinal endoscopic procedure has been designed and built. The developed slave robotic system consists of a long and flexible body that allows it to follow the endoscope through the human natural orifice. This characteristic has the potential to allow treatment such as suturing to be performed on the patient without the need of any incision. A user was able to apply the master-slave system to perform tasks such as grabbing and cutting as well as pick and place. Future animal trials are planned using the developed system.

ACKNOWLEDGMENT

This work was supported in part by the National Medical Research Council (Singapore) under Grant NMRC/0827/2004 and Nanyang Technological University (Singapore) under grant SUG 44/04.

REFERENCES

Carrozza, M.C., Dario, P., & Phee, L. (2003). Micromechatronics in surgical. *Transactions of the Institute of Measurements and Control, 25*(4), 309-327.

Intuitive Surgical Incorporation. *The da Vinci surgical system*. Retrieved February 7, 2008, from http://www.intuitivesurgical.com/

Hirose, S., & Chu, R. (1999). Development of a lightweight torque limiting M-drive actuator for hyper-redundant manipulator float arm. In *Proceedings of the 1999 IEEE International Conference on Robotics & Automation* (pp. 2831-2836).

Kaneko, M., Paetsch, W., & Tolle, H. (1992). Input-dependent stability of joint torque control of tendon-driven robot hands. *IEEE transactions on Industrial Electronics, 39*(2), 96-104.

Low, S.C., & Phee, L. (2004). A review of master-slave robotic systems for surgery. In *Proceedings of the 2004 IEEE Conference on Robotics, Automation and Mechatronics* (pp. 37-42).

Phee, L., Ng, W.S., Chen, I.M., Seow C.F., & Davies, B.L. (1997). Locomotion and steering aspects in automation of colonoscopy. *IEEE Engineering in Medicine and Biology Magazine, 16*(6), 85-96.

Medical Privacy and the National Health Information Network Initiative

Edward J. Szewczak
Canisius College, USA

INTRODUCTION

Traditionally, medical data were typically collected and stored as records in physicians' offices and hospitals. Often the data were recorded manually and retrieved manually. Today collection and retrieval of medical data are increasingly done using information technology (IT). The introduction of IT, especially e-technologies, has changed the handling of medical data in ways that are both promising for improving health care as well as threatening to the individual patient's medical privacy.

This chapter will examine medical privacy and the National Health Information Network (Kaushal et al., 2005) initiative. Other technologies such as radio frequency identification device (RFID), voice over IP (VOIP), and telemedicine, while relevant to the topic of medical privacy and IT, are tangential to the more central issues of electronic health records, medical databases, and the development of a national health information network. Although the focus will be on medical privacy in the United States, it is evident that people in other countries are also concerned about their medical privacy. Singaporeans, especially Malays, worry that their medical histories may result in racial discrimination. In the area of medical research, the Chinese and Indians want their medical information confined only to the area of study that was originally agreed to and not disseminated widely (Tan, 2006).

THE MEDICAL RECORD

Patient data form the *medical record* and its contents (www.eff.org/Privacy/Medical/1993_ota_medical_privacy.report).

Medical records may contain patient data such as name, address, age, next of kin, names of parents, date and place of birth, marital status, religion, history of military service, Social Security number, name of insurer, complaints and diagnoses, medical history, family history, previous and current treatments, an inventory of the condition of each body system, medications taken now and in the past, use of alcohol and tobacco, diagnostic tests administered, findings, reactions, and incidents. Records may also contain subjective information based on impressions and assessments by health care workers such as mental ability and psychological stability and status. In addition to data about the patient's current condition, the medical record may also contain the results of genetic research and testing that enable predictions of future medical conditions and the prospects of developing specific medical problems.

Traditionally, the creation and maintenance of medical records were done manually by health professionals, mostly as an aid to memory. But IT is changing this practice. Notes handwritten by doctors and nurses are being put into electronic form in the name of faster, more extensive access to needed information. Healtheon and other health care companies are competing to get doctors to write prescriptions over the Internet and to persuade people to place their personal health records on the Internet (Consumer Reports, 2000). Companies have made software available that an individual can use to create an Internet-based personal health record that can be used to organize family medical histories, including medical conditions, medications, and allergies. These personal records may be transmitted to health professionals over a computer network (Rubenstein, 2005).

Medical records are available online to medical practitioners for the purposes of decision-making and improving health care. They are also available to other users and institutions in nontreatment contexts. Medical records are used to conduct federal government-mandated medical community audits of physician competency and performance. They are also used by insurance companies in the assessment of an applicant's eligibility for health and life insurance and in claims processing to detect medical fraud. Medical information is also used by private employers, educational institutions, credit investigators, and law enforcement agencies for a variety of nonmedical reasons.

THE IMPORTANCE OF MEDICAL PRIVACY

Cate (1997) identified a number of conceptions of what constitutes privacy from the literature. Privacy has been viewed as an expression of one's personality or personhood, focusing on the right of the individual to define his or her essence as a human being; as autonomy—the moral freedom of the individual to engage in his or her own thoughts, actions, and decisions; as citizens' abilities to regulate information about themselves and thus control their relationships with other human beings; and as secrecy, anonymity, and solitude.

In the area of medical information, the definition of *privacy*—"the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others" (Westin, 1967, p. 7)—is consistent with the confidential relationship between doctor and patient. Confidentiality refers to how data collected for approved purposes will be maintained and used by the individual, group, or institution that collected it, what further uses will be made of them, and when individuals will be required to consent to such uses. Privacy may be construed as a balance struck by society between an individual's right to keep information confidential and the societal benefit derived from sharing the information and how the balance is codified into legislation, giving individuals the means to control personal information (Office of Technology Assessment, 1993).

Various public opinion polls conducted since 1993 have uncovered a basic concern people have about the privacy of their medical records and how these records may be used (www.epic.org/privacy/medical/polls.html). Among the major concerns of those polled are the following:

- Personal health insurance information might be used by employers to limit job opportunities.
- Sensitive health information might leak because of weak data security.
- Protecting the confidentiality of medical records is absolutely essential or very important in health care reform.
- Existing federal health privacy rules protecting patient information may be reduced in the name of efficiency.
- The trend of the health care system to keep medical records electronically raises threats to their privacy.
- Medical records are more secure on paper than in electronic form.
- Medical and government researchers should not be allowed to study an individual's genetic information unless they first obtain his or her consent.
- Government agencies, insurance companies, and researchers are allowed to see medical records without the patient's permission.
- The federal government is requiring everyone to be assigned a medical identification number, similar to a Social Security number, to create a national medical database.
- Insurance companies are getting more information from patients' doctors than is needed.
- Medical information from a computerized national health information system will be used for many nonhealth purposes.
- It is important that individuals have the legal right to obtain a copy of their own medical record.

Federal and state governments have attempted to deal with these issues in ways that satisfy the needs of various stakeholders, such as doctors, dentists, chiropractors, insurance companies, researchers, law enforcement, and data processing firms, as well as individuals. The result has been various legislative measures that provide legal compromise for the stakeholders.

MEDICAL PRIVACY LEGISLATION

It should be noted that the Bill of Rights does not address privacy issues at all. However, in *Griswold v. Connecticut* (381 U.S. 479 (1965)), the Supreme Court found sources for a right to privacy in the First, Third, Fourth, Fifth, and Ninth Amendments to the Constitution of the United States in the form of "zones" or "penumbras" of privacy (www.eff.org/Privacy/Medical/1993_ota_medical_privacy.report). A major modern discussion of an information privacy right is *Whalen v. Roe* (429 U.S. 589 (1977)), wherein the Supreme Court accepted that a right of privacy includes a generalized "right to be let alone," which includes "the individual interest in avoiding disclosure of personal matters." The court noted that it was "not unaware of the threat to privacy implicit in the accumulation of vast amounts of personal information in computerized data banks or other massive government files." However, the court has not expanded on this idea in any significant way (National Research Council, 1977).

The first comprehensive set of federal regulations of health information is provided by the Health Insurance Portability and Accountability Act (HIPAA) of 1996. HIPAA provides for two rules related directly to medical privacy: the Privacy Rule (45 *Code of Federal Regulations* 164.500–164.534) and the Security Rule (45 *Code of Federal Regulations* 164.103–164.318) (www.archives.gov/federal-register/index.html).

The HIPAA Privacy Rule

The HIPAA Privacy Rule provides the federal floor of privacy of protected health information (PHI) in the United States. It only applies to medical records maintained by covered entities (i.e., health care providers, health plans, and health care clearinghouses/data processing firms) in any form (electronic or nonelectronic, including oral). It allows more stringent state laws to continue in force. An individual has a number of rights under the Privacy Rule, including the following (adapted and expanded from www.epic.org/privacy/medical):

- To access, inspect, and copy PHI held by hospitals, clinics, health plans, and other covered entities with some exceptions.
- To request amendments to PHI held by covered entities.
- To request an accounting of disclosures that have been made without authorization to anyone other than the individual for purposes other than treatment, payment, and health care operations (i.e., medical practice evaluations for accreditation).
- To receive a Notice of Privacy Practices from doctors, hospitals, health plans, and others in the health care system.
- To request restrictions on uses and disclosures of PHI.
- To complain about privacy practices to a covered entity and to the Secretary of Health and Human Services.

Although many health care practitioners have been concerned with the privacy and confidentiality of patient information as a matter of course, HIPAA ensures that standards are being applied equally (Deshmukh & Croasdell, 2005). The Privacy Rule includes civil and criminal penalties for violations of an individual's privacy. Criminal penalties can approach $250,000 and/or 10 years imprisonment if the offense is committed with intent to sell, transfer, or use PHI for commercial or personal gain, or for malicious harm.

It should be noted that the HIPAA Privacy Rule does not prohibit the disclosure of PHI when such disclosure is required or permitted by other federal law. For example, the Gramm-Leach-Bliley Act does not prohibit the sharing of information among affiliated companies (such as banks and brokerages, which are *not* covered entities). So an individual's credit card account transactions may include data about where an individual goes for health care, and this data may be shared among affiliated companies and is not protected by HIPAA. The HIPAA Privacy Rule also explicitly includes a number of exceptions to the rules for use and disclosure. In fact, there are a number of uses and disclosures of information for which an authorization or opportunity to agree or object is *not* required (for example, for judicial and administrative proceedings, and for law enforcement purposes), including the use of PHI for marketing purposes (which, according to the Department of Health and Human Services, may be too difficult to distinguish from treatment purposes) (www.privacyrights.org/fs/fs8a-hipaa.htm).

Because HIPAA does not prohibit the sharing of PHI among various covered entities or their business associates, PHI could be used in ways other than for treatment or billing. An individual could be charged higher loan rates because of some piece of data in his or her medical record, and it would be impossible to prove the data were shared, because there is no required disclosure audit. Also, data networks may be Internet-based and global in reach. Individual health records may be transmitted overseas and handled by subcontractors in ways the individual is completely unaware. Another security challenge is the data breach. A hacker or even a trusted employee can steal data from a computer system and offer them for sale to interested parties (*Consumer Reports*, 2006).

Arguing that under HIPAA anyone can disclose, use, sharem and sell PHI without the individual's knowledge or consent, Senator Patrick Leahy proposed the Medical Information Privacy and Security Act in 1999, which provides for the following patient's rights (www.leahy.senate.gov/press/199903/990310a.html):

- To inspect, copy, and supplement your own medical record.
- To a clear explanation of who can see your health information and why before it is disclosed.
- To limit disclosure of your health information only to those directly involved with your health care.
- To have privacy protections for your medical records as strong as your video rental records by requiring law enforcement agents to first obtain a warrant before access to your health files.
- To ask your state officials for even stronger medical privacy protections.

The HIPAA Security Rule

The HIPAA Security Rule provides security standards and implementation specifications for three kinds of safeguards (administrative, physical, and technical) to protect PHI in *electronic* form. It also divides the implementation specifications into required and addressable (i.e., not required but recommended). Covered entities have a certain amount of flexibility in implementing addressable specifications. In deciding which security measures to adopt, the covered entity must consider its own size, complexity and capabilities, technical infrastructure, hardware and software security capabilities, the costs of security measures, and the probability and criticality of potential risks to electronic PHI. For example, covered entities may choose to adopt encryption as a technical safeguard for the transmission security standard. But since encryption is given as an addressable implementation specification, it is not required by HIPAA but simply recommended. In addition, while encryption and other technologies may keep patient data private and secure, it is what people who have access to the decrypted data do with it that is important to the issue of medical privacy (Patton, 2005).

HIPAA is not specific as to the exact technology that should be used to implement transmission security, since technology changes and progresses in ways that are difficult to predict. Current implementation of transmission security will most likely involve the use of firewalls, user authentication, encryption/decryption, antivirus/malware software, and anonymizers (Cheng & Hung, 2006; Fung & Paynter, 2006; Janczewski, 2000). These implementation choices will be replaced as newer, more effective technologies become available.

NATIONAL HEALTH INFORMATION NETWORK

The federal government is promoting a national system of electronic health records (EHRs) and the building of a National Health Information Network (NHIN), which will connect EHRs to health care providers, insurers, pharmacies, laboratories, and claims processors (Kaushal et al., 2005). HIPAA makes explicit mention of Electronic Data Interchange (EDI). However, implementation of EDI by covered entities has resulted in many proprietary EDI formats, resulting in a lack of common industrywide standards. This lack of uniformity is viewed as a major obstacle to realizing potential efficiency and savings (45 *Code of Federal Regulations*, Part 162). Since EDI has been replaced by the use of TCP/IP in Internet-based networks, it is likely that any future NHIN will use TCP/IP as its fundamental protocol, perhaps together with legacy systems for a time (www.amia.org/pubs/symposia/D005234.pdf ; Cheng & Hung, 2006; Deshmukh & Croasdell, 2005).

One proposal for NHIH implementation is to mimic the Santa Barbara County Data Exchange (SBCDE) on a national level. The SBCDE is a peer-to-peer network in which medical files are exchanged directly among users. A central host provides security and linking functions but does not store data. Suppliers of data must integrate data from a number of sources (e.g., pharmacies, laboratories, physicians) (Kaushal et al., 2005).

Another approach to developing an NHIN is the Veterans Administration's Veterans Health Information System and Technology Architecture (VistA), which was introduced in 1996 and is based on the client-server architecture. The VA plans to develop Health*e*Vet-VistA, which will allow veterans to access their medical records online (www.va.gov/vista_monograph/docs/vista_monograph2005_06.pdf). *Consumer Reports* (2006, p. 40) relates how the system currently works:

Two physicians at the Veterans Affairs Medical Center in Washington, D.C., showed us how a patient's computerized record gives them access to layers of information, including notes from office visits, hospital admissions and discharge notes, special patient problems, allergies, diagnostic test results, and a list of the patient's medications.

Alerts signal the doctor if the patient is due for a test or procedure. With a click, test results fill the screen. … A doctor at any of the VA's 1,300 care-center locations across the nation can pull up a patient's file and add information to it if a veteran is treated at that facility.

MEDICAL DATABASES

The role of medical databases in an NHIN is extremely important. However a database is implemented, the EHRs comprising it will be accessed by many interested parties over the NHIN. There is a number of benefits as well as disadvantages of medical databases (www.lbl.gov/Education/ELSI/privacy-main.html). Among the benefits are the following:

- A patient's medical information would be immediately available to an attending doctor, including life-saving information.
- Researchers would be able to track certain diseases as well as patients' responses to certain drugs.
- Medical databases would allow for better organization and more legibility of medical files.
- EHRs may be more secure than paper records since security systems can monitor medical databases.

Among the disadvantages of medical databases are the following:

- Employers may access medical information about their employees, which they might use to deny employment or job advancement.
- Insurers may use medical information to deny insurance to people they consider to be high risk.
- Digitizing medical records will allow many more people legitimate access to medical records, with the increased possibility that the information may be misused by one or more of them.

One of the largest central databases of EHRs is the Medical Information Bureau (MIB). It is shared by insurance companies to obtain information about life insurance and individual health insurance policy applicants. If the applicant reports a condition that the insurer considers significant, or if the results of a required examination, blood test, or urine test raise questions for the insurer, the insurer will report that information to the MIB. MIB EHRs consist of codes indicating a particular condition or lifestyle (such as the individual smokes cigarettes). As such, MIB does not include the totality of an individual's medical record (www.privacyrights.org/fs/fs8a-hipaa.htm).

The Children's Hospital of Philadelphia (CHOP) is collecting DNA profiles on as many as 100,000 child patients in order to develop an anonymous database that researchers can use to study children's genetic profiles. Research results may reveal which genes underlie problems affecting children, such as diabetes, obesity, asthma, and cancer. This research could lead to the development of diagnostic tests and drugs. By linking genetic information to EHRs, CHOP may obtain research funds and patents and forge partnerships with drug companies (Regalado, 2006).

Various federal, state, and local governments maintain databases of personal (including medical) information. As *Consumer Reports* (2000, p. 23) notes:

The federal government maintains electronic files of hundreds of millions of Medicare claims. And every state aggregates medical data on its inhabitants, including registries of births, deaths, immunizations, and communicable diseases. But most states go much further. Thirty-seven mandate collection of electronic records of every hospital discharge. Thirty-nine maintain registries of every newly diagnosed case of cancer. Most of these databases are available to any member of the public [emphasis added] who asks for them and can operate the database software required to read and manipulate them.

A computer privacy researcher at Carnegie Mellon University was able to retrieve the health records of the governor of Massachusetts from an "anonymous" database of state employee health insurance claims by knowing his birth date and ZIP code. The researcher demonstrated that she could do the same for 69% of the 54,805 registered voters on the Cambridge, Massachusetts, voting list (*Consumer Reports*, 2000).

Although many of these government database records are stripped of information that could be used to identify individuals (e.g., Social Security numbers), it is still possible to link the records to private sector medical records using standard codes for diagnoses and procedures employed by the United States health care system. The codes are usually included on insurance claims and hospital discharge records. In addition, a patient's anonymity may be compromised by the fact that personally identifiable health information is needed for a variety of research purposes (e.g., to check for duplicate records or redundant cases and for longitudinal studies) (www.epic.org/privacy/medical/GAO-medical-privacy-399.pdf).

Electronically available public records (e.g., court records) are also a source of an individual's medical information (Ogles, 2004). An individual's medical record may be entered into court documents (i.e., if an individual sues over payment claims), which are available online. Public records also have a connection to junk mail, since counties have sold information from public records to commercial companies that then repackage it and resell it to other companies and individuals (Leach, 2004). Junk mail in itself may not be overly troublesome to an individual. But what these companies and individuals may do with public record information in addition to creating and sending junk mail is cause for some concern.

Much personal health information that is available to the public is volunteered by individuals themselves by responding to 800 numbers, coupon offers, rebate offers, and Web site registration. The information is included in commercial databases like Behavior-Bank sponsored by Experian, one of the world's largest direct-mail database companies. This information is sold to clients interested in categories of health problems, such as bladder control or high cholesterol. Drug companies are also interested in the commercial databases (*Consumer Reports*, 2000). This interest will be heightened as hospitals link up electronically with doctors' offices' medication records (Landro, 2006). Data mining is often the rationale for wanting access to this information (Krzysztof & Moore, 2002; Prather, Lobach, Goodwin, Hales, Hage & Hammond, 1997). (For an account of the benefits of data mining to consumers, see Cook & Cook, 2004).

FUTURE TRENDS

The future of medical information will involve the use of an NHIN in increasingly sophisticated ways. However, the greatest obstacle to the success of NHIN implementation appears to be the medical profession itself (Patton, 2005). Many physicians have small practices with few extra resources or connections to large medical institutions. These physicians have little financial incentive to acquire expensive IT needed to link up with an NHIN. Nor do physicians want to have to pay to share medical information. Finally, many physicians do not have the time to log on to a computer to retrieve medical information. It seems that a physician champion of IT (perhaps the U.S. Surgeon General) will be needed in order to convince fellow physicians of the worth of an NHIN.

Patients will also have a say in how their medical information will be shared by the various HIPAA-covered entities. As the ultimate consumers of medical services, how patients view the information sharing practices of doctors, nurses, administrators, claims processors, and registration clerks in various hospitals and offices will impact how a NHIN is valued and used.

CONCLUSION

The implementation of am NHIN is technologically feasible and offers great promise for lowering medical costs, improving medical decision-making, and providing better quality medical care. How the users of the NHIN safeguard the medical privacy of patients will ultimately determine the success of the NHIH initiative. In particular, physicians must be convinced of the worth of the information sharing that the NHIH will make possible and feel confident that doctor-patient confidentiality will not be jeopardized. Patients must be convinced that their medical privacy will be assured and not sacrificed to commercial interests.

REFERENCES

Cate, F.H. (1997). *Privacy in the information age*. Washington, D.C.: Brookings Institution Press.

Cheng, V.S.Y., & Hung, P.C.K. (2006). Health insurance portability and accountability act (HIPAA) compliant access control model for Web services. *International Journal of Healthcare Information Systems and Informatics, 1*(1), 22–39.

Cook, J.S., & Cook, L.L. (2004). Compliance with data management laws. In L. Brennan, & V. Johnson (Eds.), *Social, ethical and policy implications of information technology* (pp. 251–272). Hershey, PA: Idea Group Inc.

Deshmukh, P., & Croasdell, D. (2005). HIPAA: Privacy and security in health care networks. In A.C. Peace, & L. Freeman (2005). *Information ethics: Privacy and intellectual property* (pp. 219–237). Hershey, PA: Information Science Publishing.

Fung, M.Y.L., & Paynter, J. (2006). The impact of information technology in healthcare privacy. In G. Yee (Ed.), *Privacy protection for e-services* (pp. 56–93). Hershey, PA: Idea Group Publishing.

Janczewski, L.J. (2000). Information security framework for health information systems. In A. Armoni (Ed.), *Healthcare information systems: Challenges of the new millennium*. Hershey, PA: Idea Group Publishing.

Kaushal, R., et al. (2005). The costs of a national health information network. *Annals of Internal Medicine, 143*(3). Retrieved from www.annals.org/cgi/reprint/143/3/165.pdf

Krzysztof, J.C., & Moore, J.C. (2002). Uniqueness of medical data mining. Retrieved from isl.cudenver.edu/Publications/Papers/UniquenessOfMedicalDataMining.pdf

Landro, L. (2006). What drugs do you take? Hospitals seek to collect better data and prevent errors. *The Wall Street Journal*, May 23, D1.

Leach, S.L. (2004). Privacy lost with the touch of a keystroke? Retrieved from www.csmonitor.com/2004/1110/p15s02-stin.html

National Research Council. (1977). *For the record: Protecting electronic health information*. National Academy Press. Retrieved from http://books.google.com/books?hl=en&id=i1IBqxG5bqYC&dq=for+the+record+protecting+electronic+health+information+national+research+council&printsec=frontcover&source=web&ots=GLIex6EVoZ&sig=TYYEVhxxrROa43wlXAs7TxTQl9A#PPP1,M1

The new threat to your medical privacy. (2006). *Consumer Reports*. March, 39–42.

Office of Technology Assessment. (1993). Protecting privacy in computerized medical information. Retrieved from www.wws.princeton.edu/ota/disk1/1993/9342/9342.pdf

Ogles, J. (2004). Court documents not fit for Web? Retrieved from www.wired.com/news/privacy/0,1848,65703,00.html

Patton, S. (2005). *Sharing data, saving lives*. Retrieved from www.cio.com/archive/030105/healthcare.html

Prather, J.C., Lobach, D.F., Goodwin, L.K., Hales, J.W., Hage, M.L. & Hammond, W.E. (1997). *Medical data mining: Knowledge discovery in a clinical data warehouse*. Retrieved from dci.mc.duke.edu/PDF_Files/Data%20Mining.pdf

Regalado, A. (2006). Plan to build children's DNA database raises concerns. *The Wall Street Journal*, June 7, B1f.

Rubenstein, S. (2005). Next step toward digitized health records. *The Wall Street Journal*, May 9, B1f.

Tan, C. (2006). Medical data privacy—The other side of the transparency and accountability coin. Retrieved from www.yawningbread.org/guest_2006/guw-104.htm .

Westin, A.F. (1967). *Privacy and freedom*. New York: Atheneum.

Who knows your medical secrets? (2000). *Consumer Reports*, August, 22–26.

KEY TERMS

EDI: Electronic data interchange.

EHR: Electronic health records.

MIB: Medical Information Bureau.

RFID: Radio Frequency Identification Device.

Telemedicine: Medical intervention provided over a distance with the aid of technology mediation.

VOIP: Voice over IP.

M

Medical Search Engines

Bob Badgett
University of Texas at San Antonio, USA

Suresh Chalasani
University of Texas at San Antonio, USA

Rajendra V. Boppana
University of Texas at San Antonio, USA

Jacqueline A. Pugh
University of Texas at San Antonio, USA

INTRODUCTION

Twenty years ago, researchers identified the difficulties physicians have answering the clinical questions that arise during medical care (Covell, Uman & Manning, 1985). Fifteen years ago, the "evidence-based medicine" movement arose (Cohen, 2001) and exhorted clinicians to use computers to look up the answers to questions as they arise during clinical care.

Fortuitously for the proponents of evidence-based medicine, the Internet matured in the 1990s (Hersh 1996 #3370), epitomized by the launching of PubMed by Al Gore in June 1997. PubMed, developed by the National Institutes of Health, provided the first free access to millions of biomedical research articles at MEDLINE.

Unfortunately, the Internet has not fulfilled its potential to assist physicians in answering clinical questions. A recent study found that physicians obtain answers to only 40% of questions that arise during patient care (Ely, Osheroff, Chambliss, Ebell & Rosenbaum, 2004); this rate is not much different than the rate of 30% reported 20 years ago (Covell et al., 1985).

BACKGROUND

This chapter focuses on information retrieval of general medical information by health care professionals. We will review research on information retrieval and medical publishing in order to describe the idealized medical search engine for health care professionals. This chapter does not address retrieval of patient-specific information, which is done by electronic medical records. This chapter does not address retrieval of general medical information by patients, although the principles are similar.

Performance parameters of the idealized medical search engine: We propose that the idealized medical search engine should provide scientifically valid and actionable answers to 95% of questions received. In addition, the total time to use the search engine, including query formation, searching, and interpretation of results, should be less than two minutes. An ideal search engine should search and include search results from all relevant sources.

The first parameter, scientifically valid answers, comes from research showing that knowledge resources vary in the validity of their content. Antman, Lau, Kupelnick, Mosteller, and Chalmers' (1992) seminal study using cumulative meta-analysis to retrospectively summarize changing medical evidence over multiple years showed that expert opinion as reflected in review articles and textbook chapters either may conflict with evidence or may delay 15 years in incorporating new evidence. Fortunately, new online hypertexts, such as UpToDate, PIER, Clinical Evidence, and others, implement systematic editorial processes such as peer review of chapters, extensive attribution of evidence with footnoting, regular and frequent updates, systematic literature surveillance or searches, grading of source evidence, and other techniques that presumably increase the validity of the content of hypertexts. We label these online resources systematic hypertexts and propose that they replace textbooks as a principle information resource.

The importance of succinct, actionable answers is demonstrated in surveys that ask physicians desirable traits of an information resource (Ely et al., 2004) and in studies of uptake of content by physicians and medical students (Beck & Bergman, 1986; Grol, Dalhuijsen, Thomas, Veld, Rutten & Mokkink, 1998). Many of the systematic hypertexts listed previously use bulleted lists and succinct recommendations rather than the traditional prose structure of printed textbooks. This attention to usability adds further justification for these hypertexts to be a principle information resource.

The importance of providing fast answers is noted in studies of the barriers to the pursuit of evidence (Dee & Blazek, 1993; Ely et al., 2002). In a study that directly observed physician behavior, an average of two minutes was spent to answer each question (Ely et al., 1999). Although our parameter of two minutes may seem stringent, Sackett and Straus (1998) found that even two minutes was a barrier to using resources. Unfortunately, breaking the two-minute barrier will require medical publishers to write more succinct and usable content. One of the most commonly used current knowledge resources, UpToDate, may require 15 minutes to read and interpret (Lucas et al., 2004). The medical outpatient visit is packed with information management and competing demands that may reduce quality of care (Redelmeier, Tan & Booth, 1998) or even prevent care (Yarnall, Pollak, Ostbye, Krause & Michener, 2003). Therefore, physicians must find answers quickly.

The need for an information resource to answer a high proportion of questions is found in studies that show belief that an answer that exists to a question is a strong determinant of whether a physician will pursue an answer (Ely et al., 2004). We arbitrarily designate 95% as the proportion of questions that should be answered by a knowledge resource in order to encourage physicians to continue using the resource.

THE IDEALIZED MEDICAL SEARCH ENGINE

Features of the Idealized Medical Search Engine Needed to Achieve Satisfactory Performance

To achieve 95% success within two minutes, we propose that the idealized medical search engine should have the following features. First, the search engine should search multiple information resources. Second, the main targets of the search engine should be high quality systematic hypertexts. Third, MEDLINE should be among the other targets of the search engine in order to answer. Fourth, the search form should be easy to use and accept natural language or Boolean results. Fifth, query expansion should use a broad thesaurus such as the Unified Medical Language System (UMLS). Sixth, links to the documents retrieved must be carefully fused into a single page of information. The order of the fused results is based on three dimensions: publication date, publication type (primary vs. secondary vs. tertiary), and publication quality.

The search engine should access multiple information resources in order to achieve completeness of search space. Studies have shown that no single information resource can answer more than 80% of questions (Alper, Stevermer, White & Ewigman, 2001; Fenton & Badgett, 2005; Lucas et al., 2004). This search engine should be a federated search that simultaneously queries multiple resources and then returns the results to the user in a single Web page. This is in contrast to the more common structure of an Internet portal that gives the user access to multiple resources, but the user must manually search the resources in sequence. We discuss next the details on how the search results should be displayed after fusing and sorting.

The principle targets of the search engine should be high quality tertiary publications. Others have noted the need to search tertiary publications first (Haynes, 2001). Research has shown that physicians can query textbooks faster than MEDLINE (Chambliss & Conley, 1996). The quicker use of textbooks is due to the more organized representation of medial knowledge in tertiary publications in contrast to the fragmented nature of the millions of citations on MEDLINE. As already noted, the ideal tertiary publications to be targets of a search are systematic hypertexts. Systematic hypertexts strive to improve validity of content by using processes such as peer review (Goodman, Berlin, Fletcher, & Fletcher, 1994), attribution (Kunst, Groot, Latthe, Latthe & Khan, 2002), frequent updating, stating the search date of the current version (Kunst et al., 2002), objective grading of underlying evidence, and structured literature searches. In addition, systematic hypertexts strive to improve usability by use of actionable and prescriptive recommendations (Ely, Osheroff, Chambliss, Ebell & Rosenbaum, 2005; Grol et al., 1998), succinct and bulleted summaries (Beck & Bergman, 1986; Ely et al., 2005), and layered content in which summary recommendations precede detailed explanations.

MEDLINE should be among the other targets of the search engine. Lucas, et al. (2004) found that UpToDate, one of the most commonly used systematic hypertexts, was helpful in only 71% of patients. A separate study found that UpToDate covered 80% of clinical questions (Fenton & Badgett, 2005). As MEDLINE is the most comprehensive repository of biomedical knowledge, it can provide relevant information for uncommon topics not covered in the derived literature. Google may provide this role as well (Greenwald, 2005) but has not been formally compared to MEDLINE. Unfortunately, MEDLINE is very difficult to search due to its size (15 million articles), difficult linguistics, and unstable nature. Although it is clear that the common method of searching MEDLINE—manually formulating a Boolean search and manually revising the strategy when too few or too many results are retrieved—consumes too much physician time, the best way to search MEDLINE is not clear. Two options are Boolean searching with automated revision of searches that retrieve excessive or too few results (used by SUMSearch) vs. relevancy searching with weighting of results utilizing PageRank (used by Google). There are theoretical reasons why SUMSearch's method might be better. Medical research findings are unstable; a third of high impact trials are subsequently refuted or attenuated (Ioannidis, 2005). Unfortunately, the initial findings may be reported in a high impact journal, whereas the subsequent refuting study may be in a lower impact journal (Ioannidis, 2005). Boolean queries allow results sorted by publication date, which might reduce the chance of a physician overlooking the second publication in a journal that has a lower PageRank; however, this has not been tested. In addition, as Google relies on links to documents to become aware of new documents, Google may not be as good at finding newly published research. These problems could become worse as the Internet ages and the need increases to recognize the dimension of time in the display of search results. Currently, the Internet is young compared to the age of MEDLINE.

The search form should be easy to use and accept natural language or Boolean results. Some physicians have difficulty with logical operators (Kingsland, Harbourt, Syed & Schuyler, 1993; Walker, McKibbon, Haynes & Ramsden, 1991). Automated query expansion, using a broad thesaurus such as the Unified Medical Language System (UMLS), serves two purposes. First, when searchers must manually browse a thesaurus, underutilization of canonical terms occurs when synonymous medical words exist (Kingsland et al., 1993). Second, use of a thesaurus also allows the search engine to vary the targets of the search based on the location of the search term in the thesaurus' tree structure. For example, if a search term is from the infectious disease branch of the thesaurus, a search of the CDC Web site can be included. If the search contains a drug name, searches of a drug resource and the FDA are included.

Links to the documents retrieved on the Web page returned to the physician must be carefully fused and sorted. When reviewing search results, clinicians have trouble screening out needless hits (Williamson, German, Weiss, Skinner & Bowes, 1989) and not recognizing incomplete searches (Walker et al., 1991).The order of the fused results are based on two dimensions: publication type (primary vs. secondary vs. tertiary) and publication date.

We proposed the following display order. Please note that this schema is theorized but has not been empirically studied. The principal resource, a systematic hypertext, is displayed first whenever it has content that matches all search terms. Next, the search engine would emphasize the display of other resources when their results are published more recently than the last update of the systematic hypertext, or when their results better match the search terms.

The procedure indicated in Table 1 creates the display order. For each search submitted to the proposed search engine, the results found from all resources are summarized in a temporary data array that is sorted according to the rules shown in Table 1.

Three scenarios to demonstrate sorting described in Table 1.

In the first scenario, the systematic hypertext has the most recent content and also matches all search terms. The search engine displays the results from the systematic hypertext first; then the engine provides the remaining links in a section titled "For more detailed information."

The second scenario is the same, except there is a randomized controlled trial that is very recently published and not covered by any of the derived literature (tertiary or secondary publications). The search engine displays the recent research first in a section labeled emerging research, then displays the systematic hypertext, and finally displays the other links in a section titled "For more detailed information."

Table 1. Dynamically determining the display order of search results

Publication Type	Example of This Publication Type	Notes	Primary Sort Order*	Secondary Sort Order
Tertiary publication	Systematic hypertext such as UpToDate or PIER		1	
Secondary publication	Practice guideline such as found at the National Guidelines Clear-inghouse	Among secondary publications, these are listed first because they usually have a broader scope and less frag-mentation than systematic reviews.	2	1
	Systematic review such as produced by the Cochrane col-laboration		2	2
	Narrative review articles	Among secondary publications, these have the least editorial quality but are very important for rapid summaries of emerging topics or for summariz-ing topics with sparse evidence.	2	3
Primary publications	Articles		3	

In the third scenario, original research is found that contains all of the user's search terms, whereas the resources in the derived literature only partially match the search terms. In this scenario, the search engine displays the original research, first in a section titled "Best matches for your search," following by the systematic hypertext and then the remaining links.

Barriers to the Implementation of the Idealized Medical Search Engine

Significant barriers hinder the development of the idealized search engine. These barriers are lack of evidence that providing rapid access to extrasomatic general medical knowledge improves patient outcomes, lack of evidence that a search engine can generate revenue for its owners, and lack of standards for B2B communications between medical publishers and search engines.

The impact of facilitating information retrieval on the quality of medical care is difficult to establish. In a cleverly designed before-and-after-study, Lucas found that 14% of medical in-patients had improved treatment plans as judged by blinded reviewers after attending physicians were provided with literature searches (Lucas et al., 2004). However, in regard to physicians doing their own searches, there is no randomized controlled trial that shows physicians with rapid access to medical knowledge provide better care than physicians without access. We acknowledge extensive literature documenting the failure to change physician behavior through education. Interventions, such as continuing medical education (CME) and practice guidelines, in which educators actively push knowledge to recipients, show physicians do not change their medical practice without intensive multifactoral interventions (Thomson O'Brien, Freemantle, Oxman, Wolf, Davis & Herrin, 2001). However, there is strong theoretical support for making knowledge accessible to physicians. Surveys show that when clinicians choose to actively pull medical knowledge to themselves by either conducting their own searches for information (Blackman, Cifu & Levinson, 2002; Brassey, Elwyn, Price & Kinnersley, 2001; Del Mar et al., 2001) or asking others to search on their behalf (Lindberg, Siegel, Rapp, Wallingford & Wilson, 1993), clinicians are very willing to let the information that they find influence their clinical care. The difference in the effectiveness of knowledge that is pulled by the clinician rather than being pushed to the clinician is consistent with theories of changing adult behavior, including that of physicians (Stross, 1999). When clinicians pull information, they have recognized a gap in their knowledge and are predisposed to changing their decision-making.

Another barrier is the lack of a business case to support a search engine. Studies of the electronic medical records (EMRs) to retrieve patient-specific medical information, which have been more thoroughly studied than retrieval of general medical knowledge, show cost savings from components of the EMR (Kaushal et al., 2005). However, the complete EMR is not financially viable, and hence, the government is considering incentives. The benefit of retrieval of general medical knowledge is not studied. There is a perception that hospitals are reluctant to invest in resources such as librarians and knowledge support (Wolf, Chastain-Warheit, Easterby-Gannett, Chayes & Long, 2002; Doyle, 2003).

The last barrier is the lack of connectivity and interoperability to support search engines to perform live searches of medical resources. Currently, very few resources offer search results in XML format. Standards such as OpenSearch (http://opensearch.a9.com/) are needed. In addition, no medical Web sites provide metadata that adequately describe the currency of the page. While Web pages display the date of the last editorial update, more important is the date of the literature search for the evidence that underlies the page's content. The Dublin Core standard (http://dublincore.org) allows for rich metadata, but the Dublin Core standard needs expanding, as the date of a Web page's literature search is not an element.

PubMed, GOOGLE, SUMsearch, AND Medscape: SEARCH ENGINES WITH CONTRASTING METHODS

We contrast four very different search engines. Each engine has many purposes aside from point-of-care retrieval of general medical information, so our comments should not be taken as summative criticisms of each engine. All four engines share the features of free access to search results and broad scope of content. Unfortunately, they also share lack of a systematic hypertext as their central target. In addition, none has adequate representation of knowledge in the display of search results. More specifically, none can vary the sorting of results by interacting the currency and publication type of the citations retrieved. The search engines are summarized in Table 2.

We now detail SUMSearch, which was created in 1998 to provide some of the features of the idealized medical search engine. In 1998, Badgett launched SUMSearch (http://SUMSearch.UTHCSA.edu), a search engine that automates a federated search for evidence from Medline and selected other resources (Booth & O'Rourke, 2000; Badgett, Paukert & Levy, 2001a). SUMSearch is unique because it (1) searches

Table 2. Comparison of selected search engines as a point-of-care tool

Search Engine	Strengths	Weaknesses
PubMed	1. Free access 2. Most comprehensive repository of world's biomedical knowledge	1. Does not have a systematic hypertext as its central target 2. Since principle clinical target of PubMed is MEDLINE, results are fragmented
SUMSearch	1. Free access 2. Live federated searching with large biomedical search space 3. Results are sorted by publication type with a secondary sort of publication date	1. Does not have a systematic hypertext as its central target 2. Results are fragmented 3. Although results are sorted on two dimensions, there is no interaction between the dimensions in order to vary the sort of publication types based on the currency of the documents retrieved
Google	1. Free access 2. Recently introduced organizing results by publication type 3. Fast	1. Does not have a systematic hypertext as its central target 2. Results are fragmented 3. Uncertain ability to search MEDLINE
Medscape	1. Free access 2. Access to full text of selected journals.	1. No identifiable primary resource in spite of Medscape's ownership of eMedicine, which has characteristics of a systematic hypertext 2. No access to important journals such as *New England Journal of Medicine* even though the Medline XML is free to Medscape 3. No logical sort and display; initial links are sponsored; other links are a fragmented collection of journal articles

multiple remote resources in real time, (2) detects when a search returns too few or too many (Blair, 1980) results and automatically revises and executes a new strategy, and (3) provides tailored suggestions to users when contingency searches cannot provide an optimal number of articles. SUMSearch has been recognized internationally (Alton & Eliasson, 2000; Anagnostelis, 2002; Booth & O'Rourke, 2000; Dearness & Tomlin, 2001; Glanville, Wilson & Richardson, 2003). Although originally developed for point-of-care searches, unexpectedly it has been used for formal literature searches for published meta-analyses. It is queried 300 to 500 times a day.

In spite of the positive published reviews already noted, positive feedback in the medical student course in which SUMSearch is used and positive e-mailed comments from the international community in a controlled trial showed that use of SUMSearch did not increase medical students' self-reports of frequency of searching (Badgett, Paukert & Levy, 2001b). We hypothesize that even when a search engine successfully delivers the relevant studies, the clinician still faces the task of synthesizing an answer by reading and integrating multiple articles during patient care. The difficulty in assimilating multiple pieces of fragmented information has been noted before (Ely et al., 2002). This hypothesis supports the need to incorporate a systematic hypertext within the metasearch and is consistent with the 4S strategy proposed by Haynes (2001).

In summary, none of the search engines adequately meet the need of the busy physician. Google and SUMSearch have the largest search space and put the most effort into how results are displayed and sorted, but both are hindered by not having a systematic hypertext as the central target of the search.

FUTURE TRENDS

Hopefully, the near future will see resolution of the barriers—financing and standards of interoperability—to the creation of a search engine. In addition, more research is needed on the best manner of searching MEDLINE.

Unresolved issues in medical publishing will have a great impact on information retrieval. It is presently unclear whether the national push for open access to medical publications will influence medical information retrieval by health care professionals during clinical care. So far, the push has been for open access to federally sponsored research. This means a large portion of original research reports (primary publications), a small proportion of federally funded meta-analyses (secondary publications), and no tertiary publications. Because the tertiary publications followed by secondary publications will be the most viewed results provided by the search engine, we doubt that the current movement toward open access to federally sponsored research will impact information retrieval. The exception is if Wikipedia's medical content matures and Wikipedia becomes an important provider of medical information.

Another issue in medical publishing is improving the readability of medical writing. As previously noted, 15 minutes may be needed to read and interpret UpToDate when answering a clinical question (Lucas et al., 2004). This time is too long. Medical writing that is for the busy physician must abandon long blocks of prose in favor of more succinct forms of text (Ely et al., 2005).

It is unclear what the NIH's role will be in providing point-of-care medical information. It has had rudimentary efforts in this area (http://gateway.nlm.nih.gov/). However, because the NLM is focused on access to content rather than creating original content, the NIH will not have a major role with its current policy. In addition, the NLM focuses more on reports of research than the hypertexts that are needed by physicians. The NLM does host books at their Bookshelf, but these are books whose authors and publishers have agreed to give open access to a copy of the book. Since high-quality tertiary publications are expensive to produce, the NIH will not have a significant role unless either Wikipedia emerges as a strong medical resource or the NLM allows publishers to retain a subscription-based business model with the NLM providing users with content.

Regarding PubMed's role as a point-of-care tool, PubMed does offer clinical queries. But as PubMed uses Boolean searching but does not automate revisions of searches that retrieve excessive results, PubMed will have a limited role in information retrieval for physicians. Research has shown how infrequently PubMed is used in this role (Ely et al., 1999).

CONCLUSION

In spite of the arrival of the Internet, retrieval of general medical information has significant problems, and 20% of office visits to primary care physicians contain questions posed by the physician that are not answered. The Internet provides the technology to create a high performance medical search engine; however, we must solve several barriers. These barriers include the justification and financing of information retrieval, improved interoperability among medical Web sites, and improvements in medical publishing that facilitate quick comprehension of medical writing.

ACKNOWLEDGMENT

Drs. Badgett and Pugh's research is supported in part by the VERDICT Research Enhancement Award Program, South Texas Veterans Health Care System Department of Medicine, and the University of Texas Health Sciences Center. The views expressed in this chapter are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs. Dr. Boppana's research has been supported by National Science Foundation (NSF) grants EIA-0117255 and ANI-0228927 and the San Antonio Life Sciences Institute grant 10001642. This research is also sponsored in part by a summer research grant to Suresh Chalasani from the University of Wisconsin system.

REFERENCES

Alper, B.S., Stevermer, J.J., White, D.S., & Ewigman, B.G. (2001). Answering family physicians' clinical questions using electronic medical databases. *J Fam Pract, 50*(11), 960–965.

Alton, V., & Eliasson, M. (2000). It's easy to search medical information on the Internet. How to search—step by step. *Lakartidningen, 97*(40), 4426–4432, 4434.

Anagnostelis, B. (2002). A guide to healthcare resources on the Internet. *Health Information & Libraries Journal, 19*(1), 59–60.

Antman, E.M., Lau, J., Kupelnick, B., Mosteller, F., & Chalmers, T.C. (1992). A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. *JAMA, 268*(2), 240–248.

Badgett, R.G., Paukert, J.L., & Levy, L. (2001a). The evolution of SUMsearch for teaching clinical informatics to third-year medical students. *Acad Med, 76*(5), 541.

Badgett, R.G., Paukert, J.L., & Levy, L.S. (2001b). Teaching clinical informatics to third-year medical students: Negative results from two controlled trials. *BMC Med Educ, 1*(1), 3.

Beck, A.L., & Bergman, D.A. (1986). Using structured medical information to improve students' problem-solving performance. *J Med Educ, 61*(9 Pt 1), 749–756.

Blackman, D., Cifu, A., & Levinson, W. (2002). Can an electronic database help busy physicians answer clinical questions? [Abstract]. *J Gen Intern Med, 17*(Suppl 1), 220.

Blair, D. (1980). Searching biases in large interactive document retrieval systems. *J Am Soc Inf Sci, 31*, 271–277.

Booth, A., & O'Rourke, A. (2000). SUMSearch and PubMed: 2 Internet-based evidence-based medicine tools. *ACP Journal Club, 132*, A16.

Brassey, J., Elwyn, G., Price, C., & Kinnersley, P. (2001). Just in time information for clinicians: A questionnaire evaluation of the ATTRACT project. *BMJ, 322*(7285), 529–530.

Chambliss, M.L., & Conley, J. (1996). Answering clinical questions. *J Fam Pract, 43*(2), 140–144.

Cohen, J.S. (2001). Dose discrepancies between the physicians' desk reference and the medical literature, and their possible role in the high incidence of dose-related adverse drug events. *Arch Intern Med, 161*(7), 957–964.

Covell, D.G., Uman, G.C., & Manning, P.R. (1985). Information needs in office practice: Are they being met? *Ann Intern Med, 103*(4), 596–599.

Dearness, K.L., & Tomlin, A. (2001). Development of the national electronic library for mental health: Providing evidence-based information for all. *Health Information & Libraries Journal, 18*(3), 167–174.

Dee, C., & Blazek, R. (1993). Information needs of the rural physician: A descriptive study. *Bull Med Libr Assoc, 81*(3), 259–264.

Del Mar, C.B., et al. (2001). Feasibility of an evidence-based literature search service for general practitioners. *Med J Aust, 175*(3), 134–137.

Doyle, J.D. (2003). A job with a view: Perspectives from the corporate side of the hospital. *J Med Libr Assoc, 91*(1), 12–17.

Ely, J.W., Osheroff, J.A., Chambliss, M.L., Ebell, M.H., & Rosenbaum, M.E. (2005). Answering physicians' clinical questions: Obstacles and potential solutions. *J Am Med Inform Assoc, 12*(2), 217–224.

Ely, J.W., et al. (1999). Analysis of questions asked by family doctors regarding patient care. *BMJ, 319*(7206), 358–361.

Ely, J.W., et al. (2002). Obstacles to answering doctors' questions about patient care with evidence: Qualitative study. *BMJ, 324*(7339), 710.

Ely, J.W., Osheroff, J.A., Chambliss, M.L., Ebell, M.H., & Rosenbaum, M.E. (2004). Answering physicians' clinical questions: Obstacles and potential solutions. *J Am Med Inform Assoc*, M1608.

Fenton, S.H., & Badgett, R. (2005). *Are there differences in online resources for answering primary care questions?* Proceedings of the AMIA Annual Symposium.

Glanville, J., Wilson, P., & Richardson, R. (2003). Accessing the online evidence: A guide to key sources of research information on clinical and cost effectiveness. *Qual Saf Health Care, 12*(3), 229–231.

Goodman, S.N., Berlin, J., Fletcher, S.W., & Fletcher, R.H. (1994). Manuscript quality before and after peer review and editing at *Annals of Internal Medicine*. *Ann Intern Med, 121*(1), 11–21.

Greenwald, R. (2005). … And a diagnostic test was performed. *N Engl J Med, 353*(19), 2089–2090.

Grol, R., Dalhuijsen, J., Thomas, S., Veld, C., Rutten, G., & Mokkink, H. (1998). Attributes of clinical guidelines that influence use of guidelines in general practice: Observational study. *BMJ, 317*(7162), 858–861.

Haynes, R.B. (2001). Of studies, syntheses, synopses, and systems: The "4S" evolution of services for finding current best evidence. *ACP J Club, 134*(2), A11–13.

Ioannidis, J.P. (2005). Contradicted and initially stronger effects in highly cited clinical research. *JAMA, 294*(2), 218–228.

Kaushal, R., et al. (2005). The costs of a national health information network. *Ann Intern Med, 143*(3), 165–173.

Kingsland, L.C., Harbourt, A.M., Syed, E.J., & Schuyler, P.L. (1993). Coach: Applying UMLS knowledge sources in an expert searcher environment. *Bull Med Libr Assoc, 81*(2), 178–183.

Kunst, H., Groot, D., Latthe, P.M., Latthe, M., & Khan, K.S. (2002). Accuracy of information on apparently credible Websites: Survey of five common health topics. *BMJ, 324*(7337), 581–582.

Lindberg, D.A., Siegel, E.R., Rapp, B.A., Wallingford, K.T., & Wilson, S.R. (1993). Use of MEDLINE by physicians for clinical problem solving. *JAMA, 269*(24), 3124–3129.

Lucas, B.P., et al. (2004). The impact of evidence on physicians' inpatient treatment decisions. *Journal of General Internal Medicine, 19*(5p1), 402–409.

Redelmeier, D.A., Tan, S.H., & Booth, G.L. (1998). The treatment of unrelated disorders in patients with chronic medical diseases. *N Engl J Med, 338*(21), 1516–1520.

Sackett, D.L., & Straus, S.E. (1998). Finding and applying evidence during clinical rounds: The "evidence cart." *JAMA, 280*(15), 1336–1338.

Stross, J.K. (1999). Guidelines have their limits. *Ann Intern Med, 131*(4), 304–306.

Thomson O'Brien, M.A., Freemantle, N., Oxman, A.D., Wolf, F., Davis, D.A., & Herrin, J. (2001). Continuing education meetings and workshops: Effects on professional practice and health care outcomes. *Cochrane Database Syst Rev,* (2), CD003030.

Walker, C.J., McKibbon, K.A., Haynes, R.B., & Ramsden, M.F. (1991). Problems encountered by clinical end users of MEDLINE and GRATEFUL MED. *Bull Med Libr Assoc, 79*(1), 67–69.

Williamson, J.W., German, P.S., Weiss, R., Skinner, E.A., & Bowes, F. (1989). Health science information management and continuing education of physicians. A survey of U.S. primary care practitioners and their opinion leaders. *Ann Intern Med, 110*(2), 151–160.

Wolf, D.G., Chastain-Warheit, C.C., Easterby-Gannett, S., Chayes, M.C., & Long, B.A. (2002). Hospital librarianship in the United States: At the crossroads. *J Med Libr Assoc, 90*(1), 38–48.

Yarnall, K.S., Pollak, K.I., Ostbye, T., Krause, K.M., & Michener, J.L. (2003). Primary care: Is there enough time for prevention? *Am J Public Health, 93*(4), 635–641.

KEY TERMS

MEDLINE: The database of citations and links to much of the world's biomedical research. MEDLINE is maintained by the National Library of Medicine of the National Institutes of Health. More information is available at http://www.nlm.nih.gov/databases/databases_medline.html.

MeSH Terms: MeSH stands for Medical Subject Headings (MeSH®). It is a set of keywords and vocabulary produced by the National Library of Medicine and used for indexing, cataloging, and searching for biomedical and health-related information and documents. More information is available at http://www.nlm.nih.gov/mesh/.

Primary, Secondary, and Tertiary Publications: Primacy publications are source documents such as journal articles and press releases. Secondary publications are summaries and syntheses of primary publications. Examples of secondary publications include practice guidelines, review articles, and meta-analyses. Tertiary publications are printed textbooks and online hypertexts.

PubMed: The public search engine for MEDLINE that is maintained by the National Center for Biotechnology Information. PubMed is available at http://pubmed.gov.

SUMSearch: SUMSearch is a federated medical search engine that in response to a query from a user performs live searches of remote Internet sites and automates revising of searches that retrieve excessive or too few documents. SUMSearch is available at http://sumsearch.uthscsa.edu.

Systematic Hypertexts: We introduce this term to distinguish online hypertexts that have a systematic method to their creation and maintenance. These methods have either been shown in studies or requested by physicians in surveys to improve the validity or usability of medical information. Examples of methods to increase validity include systematic editorial processes such as use of peer review of chapters, extensive attribution of evidence with footnoting, regular and frequent updates, systematic literature surveillance or searches, and objective grading of source evidence. Examples of techniques to improve usability include succinct, actionable answers and a standardized structure for each chapter. We propose that systematic hypertexts replace textbooks as a principle information resource.

The Unified Medical Language System (UMLS®): Metathesaurus is a vocabulary that links together multiple medical vocabularies. More information is available at http://www.nlm.nih.gov/pubs/factsheets/umls.html.

Minimal Invasive Surgery and Therapy (MIS and MIT) in Hirschsprung's Patients

Nilmini Wickramasinghe
Illinois Institute of Technology, USA

Giuseppe Turchetti
Scuola Superiore Sant'Anna, Italy

Barbara Labella
Scuola Superiore Sant'Anna, Italy

Andrea Moglia
Scuola Superiore Sant'Anna, Italy

Arianna Menciassi
Scuola Superiore Sant'Anna, Italy

Paolo Dario
Scuola Superiore Sant'Anna, Italy

INTRODUCTION

The generally accepted treatment for Hirschsprung's disease is through surgical intervention. Currently, the recognized techniques include the Swenson technique, which is less frequently used as the operation is delicate and can give rise to complications; the Duhamel technique, which is arguably the simplest approach and consistently provides good results; and the Soave technique, which also provides good results but is often more complex than the Duhamel approach. However, surgical and diagnostic procedures of the future will evolve from embracing current technologies that enable minimally invasive approaches to extremely targeted, localized, and high-precision endoluminal techniques. This requires entirely new types of surgical tools capable of entering the human body through natural orifices (by insertion, ingestion, or inhalation), very small incisions (injections), or even through skin absorption and maybe configuring themselves in complex kinetic structures at the specific site of intervention. Moreover, such approaches necessitate modification of classic surgical techniques. While the advantages of minimally invasive surgery and minimally invasive therapy (MIS and MIT) are widely acknowledged, this chapter serves to highlight the advantages of such procedures in pediatric medicine, both at the diagnostic and intervention levels and the consequent implications to classic surgical techniques. The particular focus of the chapter pertains to the specific advantages the following techniques can bring into pediatric diagnostic and surgical techniques in the case of Hirschsprung's disease: (1) endoluminal miniaturized tools for gastrointestinal endoscopy, (2) gastrointestinal capsules for digestion (e.g., M2A capsule), and (3) laparoscopic tools for surgery. Both technological and economic perspectives are discussed.

BACKGROUND

Hirschsprung's disease (also known as congenital megacolon or congenital intestinal aganglionosis) occurs when some of the nerve cells that are normally present in the intestine do not form properly while a baby is developing during pregnancy. In children with Hirschsprung's disease, a lack of nerve cells in part of the intestine interrupts the signal from the brain and prevents peristalsis in that segment of the intestine. Because stool cannot move forward normally, the intestine can become partially or completely obstructed (blocked) and begins to expand to a larger than normal size. Today it is agreed that Hirschsprung's disease can only be cured through surgery (Rehbein, Morger, Kundert & Meher-Ruge, 1966; Soper & Figueroa, 1971; San Luis, Nemoto & Beardmore, 1968; Soaue, 1964)). The goals of the surgical treatment are to establish regular and spontaneous bowl function and maintain complete continence.

The three recognized surgical techniques include (1) the Swenson technique, developed by Swenson in 1947 but today is less frequently used, as the operation is delicate and can give rise to complications; (2) the Duhamel technique, which is arguably the simplest approach and consistently provides good results; and (3) the Soave technique, which also provides good results but is often more complex than the Duhamel approach.

With the advent of technologies, minimally invasive surgery (MIS) is becoming more and more prevalent in various areas of surgery. Essentially, MIS is done through small incisions using specialized techniques, including miniature cameras with microscopes, tiny fiber-optic flashlights, and high-definition monitors, so that surgeons in many specialties can perform surgery through an incision that requires only a stitch or two to close. For patients, MIS has several benefits since it can minimize pain, speed up recovery, and eliminate potential complications. For health care delivery, MIS represents an avenue to offer high-quality health care delivery and yet minimize costs associated with longer hospital stays and recovery. The area of pediatric surgery has only just begun to investigate the possibilities of incorporating MIS techniques. We discuss the role of MIS and MIT in the context of Hirschsprung's disease.

MIS/MIT TECHNIQUES AND TOOLS

MIS and MIT are growing and evolving areas that use many tools and techniques. Depending on the specific treatment domain (e.g., spine surgery vs. gastrointestinal surgery) naturally different tools and techniques are more suitable. Three MIS/MIT techniques that are particularly suitable in the context of pediatric gastrointestinal treatment include:

1. Endoluminal miniaturized tools for gastrointestinal endoscopy
2. Gastrointestinal capsules for digestion
3. Laparoscopic tools for surgery

We briefly discuss each in turn.

ENDOLUMINAL MINIATURIZED TOOLS FOR GASTROINTESTINAL ENDOSCOPY

Pediatric gastroenterology, which was born in the early 1970s, is nowadays increasingly related to the application of gastrointestinal endoscopy for the diagnosis and treatment of digestive diseases in children (Gilger, 2001; Olives et al., 2004).

Today, gastrointestinal endoscopy plays an important role in the diagnosis of manifold pathologies, including Hirschsprung's disease (HD), which may be detected by three tests: contrast enema X-rays, manometry, and biopsy (North American Society for Pediatric Gastroenterology Hepatology, and Nutrition). In particular, rectal suction has recently demonstrated to be the most accurate test for diagnosing HD (De Lorijn et al., 2005). This procedure enables preoperative hystochemistry on mucosal-submucosal specimens for proper enteric nervous system (ENS) evaluation (Martucciello et al., 2005). It is traditionally performed using the biopsy tool conceived by Noblett (Campbell & Noblett, 1969; Noblett, 1969) in the late 1960s. By using this instrument, a physician can create suction within the lumen and then use a cutter for mucosa sampling. Although such a system is atraumatic and easy to use, it requires two operators, while the size of the specimen and the depth of sampling are inconstant. In order to overcome all these limitations, Prato, Martucciello, and Jasonni (2001) have devised the "Solo-RBT." More recently, Bio-Optica has developed a kit for enzymatic-histochemistry.

Although barium enema is not essential to confirm the diagnosis of HD in many cases, it is useful in evaluating the level of aganglionosis and aids in the decision regarding the surgical approach (i.e., transanal, transabdominal laparoscopic, or open). Manometry, although not necessary, is generally exploited with rectal biopsy and barium (Martucciello et al., 2005).

It is assumed that the development of biopsy instrumentation could benefit by both the advancements in microfabrication technologies and the employment of new and biocompatible materials. For example, by exploiting Wired Electrical Discharge Machining (WEDM) and Nitinol, miniaturized biopsy tools for a gentle approach to the GI tract are feasible (Menciassi, Moglia, Gorini, Pernorio, Stefanini & Dario (2005). Biopsy could be eventually customized for usage in pediatrics, where minimization of intrusion is critical due to the narrow spaces. Nonetheless, biopsy can be somewhat painful for patients, since it requires the removal of samples of tissue. In order to preserve the tissue, there is a growing interest to find new and less traumatic modalities to perform biopsy. In this sense, a promising technique is optical biopsy, given that the optical properties of a biological tissue depend on its biochemical composition and its cellular and subcellular structure. Therefore, optical measurements have a strong potential for the development of noninvasive *in vivo* medical diagnostic tools. Such techniques should significantly improve the efficiency of biopsies or help in determining pathologies margins in a surgical field. Several types of optical properties can be measured, such as fluorescence, Raman scattering, and absorption and scattering, each of them giving complementary information on the tissue. As an example, past studies have proved that the absorption and scattering of tumor tissues differ from their surrounding normal tissue (Bevilacqua, Piguet, Marquet, Gross, Tromberg & Depeursinge, 1999). At the Ecole Polytechnique Fédérale de Lausanne (EPFL) in Switzerland, a research group reported *in vivo* clinical measurements performed endoscopically on a human GI tract (Charvet et al., 2002; Charvet et al., 2004).

GASTROINTESTINAL CAPSULES FOR DIGESTION

The Wireless Capsule Endoscopy (WCE) indicates a swallowable and disposable vitamin-size pill able to acquire images from inside the gastrointestinal (GI) tract. This device offered for the first time a noninvasive way to enter the small intestine, which has always been known as a "black box" because of the intrinsic difficulties to enter it from either the mouth or the anus. The feasibility of WCE has been certainly possible thanks to the advancements in several technologies, including Complementary Metal Oxide Semiconductor (CMOS), Application Specific Integrated Circuit (ASIC) transmitter, and synchronous switching of Light Emitting Diode (LED).

Although interest in WCE emerged in the past few years, actually its roots date back to the middle of the 20th century when Jacobsen et al (Nebeker, 2002) developed the first device for wireless monitoring systems: endoradiosondes, also called radio-pills, for medical and biological studies. Later, Nagumo et al (1962) conceived a passive echo capsule for temperature and pH monitoring.

Today, the most famous WCE is the PillCam™ SB, announced in 2000 and currently marketed by Given Imaging, an Israeli company (Given Imaging; Moglia, Menciassi, Schurr & Dario, 2006). The PillCam™ SB Capsule Endoscope is carried through the GI tract by peristalsis and, as its name implies, works inside the small bowel. It is 26mm in length and 11mm in diameter; it is battery powered and possesses a CMOS camera able to acquire two images per second. Once captured, pictures are transmitted via an array of sensors secured to the abdomen, to the DataRecorder™ affixed to a belt worn around the patient's waist. The examination lasts eight hours and can be carried out while a patient continues normal daily activities. Later, the patient returns the DataRecorder™ for processing on the RAPID® workstation. The RAPID® application enables the physician to view and analyze the Patient Rapid Report (PRR).

Since its adoption in clinical studies, WCE has been exploited in the diagnosis of a larger and larger number of pathologies. First applications have concerned obscure gastrointestinal bleeding (OGIB) and Crohn's disease (CD), while currently WCE is emerging for celiac disease diagnosis.

In October 2003, Given Imaging got clearance to market capsule endoscopy for use in pediatric patients ages 10 to 18. The company offers a Pediatric Accessory Kit for children and small stature patients, consisting of the Pediatric Recorder Belt and the Pediatric Sensor Array. Aabakken, Scholz, Ostensen, Emblem, and Jermstad (2003) proved the feasibility of WCE in children, while Guilhon de Araujo Sant'Anna, Dubois, Miron, and Seidman (2005) performed the first clinical study. In a clinical trial, PillCam™ SB detected linear and aphthous ulcers of the jejunum or ileum not recognized by other imaging techniques (Seidman, Sant'Anna & Dirks, 2004). Additionally, WCE detected the presence of OGIB in a 2.5-year-old patient, so far the youngest one who has received WCE treatment (Kavin, Berman, Martin, Feldman & Forsey-Koukol, 2006). Moreover, WCE was evaluated in the graft vs. host disease in young patients (Silbermintz, Sahdev, Moy, Vlachos, Lipton & Levine, 2006).

The higher diagnostic yield of WCE over conventional procedures and its high degree of tolerance have suggested to expand the application of WCE to the surveillance of oesophageal disorders such as gastroesophageal reflux disease. In 2004, PillCam™ ESO was launched by Given Imaging. This pill has the same dimensions of PillCam™ SB. However, it comes with two cameras, at the front and back end of its body.

A dedicated WCE for colon, named PillCam™ COLON, was unveiled at the United European Gastroenterology Week (UEGW) 2006, Berlin, Germany. This two-camera pill is a bit larger than those for small bowel and oesophagus; in fact, it is 32mm in length and 11mm in diameter. At present, the device is undergoing first clinical tests.

LAPAROSCOPIC TOOLS FOR SURGERY

The term minimally invasive surgery (MIS) was introduced by John Wickam to describe the emerging therapeutic approach designed to minimize the traumatic insult to the patient by surgical procedures.

The evolution of MIS techniques for adults has been among the most important surgical developments of the last century. In fact, by reducing the size of the incision, the postoperative pain should be limited (Zitsman, 2003; Georgeson, 2003).

However, the adoption of MIS techniques in young patients has been a slower process than in adults for several reasons. First, surgical instruments developed for adults are cumbersome and unsafe when used in infants and children. Additionally, no single procedure propelled the evolution of MIS in neonates like laparoscopic cholecystectomy in adults. Then, the absolute length of an incision in a neonate is short and thought to be less traumatic. Yet from a relative perspective, these incisions are at least as long as the incisions for similar surgical procedures in adults.

During the past 15 years, laparoscopic surgery has been a major driving force for the technological advances that can be grouped into three categories: facilitative, enabling, and additive. *Facilitative technologies* improve the efficiency of performance of the procedure and reduce the level of difficulty of its execution, although not necessarily its safety. Examples of facilitative technologies include the newer high-power ultrasonic dissection systems (Ultracision and Autosonix), hand-assist devices, and the cutting impedance-controlled bipolar coagulation (LigaSure). *Enabling technologies* make possible certain surgical procedures that would be impossible or very difficult without them. Enabling technologies are, of course, relative to specific interventions. Examples include radio frequency (RF) thermal and other physical tissue ablators used for laparoscopic in situ ablation of hepatic tumors, inflatable bands for obesity surgery, and linear cutting endostaplers for colorectal and upper GI surgery. *Additive technologies* bring sophistication to surgical manipulations but are not considered essential, although they may, as the technology advances, replace existing systems, but this seems unlikely in the near future. Additive technologies are usually expensive and, hence, limited to only a few centers. The best example is provided by the DaVinci system (Cuschieri, 2005).

The management of HD disease has been radically changed by endoscopic surgery. In particular, the laparoscopic-assisted transanal endorectal pull-through allows early biopsies to determine the extent of aganglionic and dysfunctional bowel (Georgeson & Robertson, 2004). Mazzotti and Langer (2001) reported the first study on laparoscopic surgery to obtain full-thickness intestinal biopsies in children. They especially pointed out that laparoscopic full-thickness intestinal biopsy is safe and effective for a variety of gastrointestinal disorders in children such as HD. Ruckauer and von Dobschuetz (2005) proved the effectiveness of laparoscopically assisted colorectal resection in HD. Moreover, by using laparoscopy prior to the transanal endorectal dissection, the transition zone can be gauged precisely (Georgeson & Robertson, 2004). A new and effective technique named laparoscopic cecostomy button placement for the treatment of CD was documented by Yagmurlu, Harmon, and Georgeson (2006).

In the future, robotic capsules integrating imaging system, instruments, and suture modules could be employed in pediatrics to help physicians manage disorders that may occur in the GI tract of young patients. According to this scenario, the surgeon will be able to guide the robot into the exact positions where lesions may be present, by means of a dedicated computerized console. In particular, robotics will bring several advantages to pediatrics surgery. First and foremost, the associated *tremors* to the surgeon's hands during their movement, crucial when working in narrow spaces, could be countered by the computer interface of the robot. Moreover, with *motion scaling*, the computer will be able to scale down a movement of several centimeters of the surgeon's hand into a small fraction. The scale factor may be varied and be independent from translations and rotations of the tiny surgical instrumentation. Another benefit of robotics is *indexing*, the ability to continue moving a tool continuously in any direction beyond the limits of the surgeon's arm (Hanly & Talamini, 2004; Kant, Klein & Langenburg, 2004; Woo, Le, Krummel & Albanese, 2004).

SOME ECONOMIC CONSIDERATIONS

Even if innovation in pediatrics, more than in other medical and surgical fields, is often driven by ethical considerations rather than by its economic implications (Goodman, 2004), some economic notes about the diffusion and cost-effectiveness of the herein proposed technologies could be usefully introduced in order to contribute to the debate—still controversial in literature—of technological innovations as the culprit of increasing costs in health care.

The diffusion of new technology in pediatric practices is now quite broad. A study carried out in 1998 by of the University of Arkansas School for Medical Sciences and Arkansas Children's Hospital resulted in 82% of the surveyed pediatric surgeons (all the members of the American Pediatric Surgical Association were surveyed) performing MIS. Laparoscopic cholecystectomy was the procedure more frequently reported (Christensen, Anthony & Roth, 2004). The results of this study suggest that the diffusion of new technology in pediatric surgery could follow the same pathway as for adult general surgery.

It is worth remarking that the benefits of MIS in children could appear (and so it was at the beginning of its application in pediatrics) less evident than in adult surgery: as children's bodies are smaller than adult bodies, surgical incisions are naturally smaller; children naturally recover in shorter time than adults; and anesthetic and analgesic doses are naturally less than for adults (Cataliotti, 2003). Even costs of minimally invasive pediatrics procedures could appear higher than for conventional surgical treatments: MIS procedures could take more time, which would entail increasing costs of medical and nursing staff and higher costs for using the operating rooms; furthermore, disposable equipments have to be used (Cataliotti. 2003). Nevertheless, regarding Hirschsprung's disease, laparoscopically assisted primary endorectal pull-through is now considered the standard treatment (Firilas, Jackson & Smith, 1998).

While approaching new technology in pediatrics from a Health Technology Assessment (HTA) perspective and focusing first on the potential cost-effectiveness of MIS compared with conventional treatments, the following elements should be considered:

- Patient outcomes
- Operative times
- Anesthetic requirements
- In-patient hospital days
- Complications
- Length of follow-up
- Pain
- Cosmetic results
- Psychological impact

Evidences already exist supporting the cost-effectiveness of new, minimally invasive treatments for Hirschsprung's disease.

To give an example, the study carried out at the Section of Pediatric Surgery of University of Tennessee during the period of February 1993 and March 1996 could be mentioned (Zitsman, 2003). Data were collected on 33 children suffering from Hirschsprung's disease; 18 of them were treated through the traditional (open) two-stage Duhamel procedure, while on 15 of them, the single-stage laparoscopic pull-through was performed.

The operative times were not considerably different for the two procedures, while the number of in-patient hospital days was significantly shorter for patients treated through laparoscopic rather than open procedure ($2{,}5 \pm 3{,}5$ vs. $5{,}3 \pm 1{,}9$ days). Complications were more frequent in the group treated with open technique that often required further interventions. The length of follow-up was longer in the open group ($12{,}2 \pm 14{,}4$ months vs. $6{,}7 \pm 3{,}2$ months). The total cost of managing Hirschsprung's disease through minimally invasive procedure was $19,088 ± 13,075, while for the open intervention, the total cost was $ 34,102 ± 19,443.

If approaching Hirschsprung's disease from a cost of illness perspective, intangible costs (and intangible cost savings) should also be considered. So less pain, better cosmetic results, and consequent less negative psychological impact would increase the advantages of minimally invasive treatment for Hirschsprung's.

The debate about cost-benefits/effectiveness of new technology in pediatrics is still very open, but as for health care technology in general, we would stress two points that we consider very critical while assessing innovative technologies:

1. The "global approach": to take into consideration and evaluate all aspects and implications of new technology, such as:
   - Technical properties
   - Safety
   - Efficacy and/or effectiveness
   - Economic impact
   - Social, legal, ethical, and/or political impacts
2. The "early assessment": new health technologies need to be carefully assessed from the first phases of their development process, even if no scientific evidence is still available. Currently, HTA methods are only applied at the end of the biomedical products development process, usually at the commercialization phase, with the aim to support the decision makers in their decision concerning the introduction of new health technologies.

   We believe that each stage of the health technology development process, from concept to final outcome, should be followed by HTA analyses carried out through models, simulations, and evaluation protocols in order to provide information about potential costs, benefits, effects, and impacts to technologists, producers, and users. This could also help researchers, developers, and producers adjust the technology development trajectory toward more effective, safe, and/or economic standards.

DISCUSSION AND CONCLUSION

The primary purpose of this chapter is to highlight the possibility for incorporating MIS/MIT tools and techniques into pediatric surgery. We did this by first discussing the instance of Hirschsprung's disease, a problem experienced by children who today are treated through complex surgical techniques that may have unpleasant side-effects due to the nature of the surgery that takes time for the patient to recover fully. We have suggested that a combination of three MIS/MIT tools and techniques should be considered in this surgical procedure. In doing so, we believe that it is indeed possible to address the orthogonal goals within health care delivery of cost-effective quality delivery of care. While we have no empirical data at this point to substantiate our claims based on other areas where MIS/MIT tools and techniques have enabled superior results, we are confident that the incorporation of these techniques into the surgical treatment of Hirschsprung's disease will also enjoy similar success.

In closing, we note that the implications of this chapter are somewhat deeper and more generalizable. As globally health care delivery struggles to satisfy the orthogonal goals of cost-effective and quality health care delivery, we can see that in surgery (as well as other treatment/clinical initiatives) and administratively, the incorporation of technology is critical. It becomes imperative on researchers and health care professionals then to continuously look for ways to incorporate technology so that superior health care delivery can ensue.

REFERENCES

Aabakken, L., Scholz, T., Ostensen, A.B., Emblem, R., & Jermstad, T. (2003). Capsule endoscopy is feasible in small children. *Endoscopy*, 798.

Bevilacqua, F., Piguet, D., Marquet, P., Gross, P.D., Tromberg, B.J., & Depeursinge, C. (1999). In vivo local determination of tissue optical properties: Applications to human brain. *Appl Opt, 38*(22), 4939–4950.

Bio-Optica Web site. http:// www.bio-optica.it

Bufo, A.J., Chen, M.K., Shah, R., Gross, G., Cyr, N., & Lobe, T.E. (1999). Analysis of the costs of surgery for Hirschsprung's disease: One stage laparoscopic pull-through versus two-stage Duhamel procedure. *Clinical Pediatrics, 38*, 593–596.

Campbell, P.E., & Noblett, H.R. (1969). Experience with rectal suction biopsy in the diagnosis of Hirschsprung's disease. *J Pediatr Surg, 4*, 410–415.

Cataliotti, F. (2003). Ethics and technique: Can we play "balloon game" with them? The case of conventional surgery versus innovative techniques. *Italian Journal of Pediatrics, 29*, 90–92.

Charvet, I., et al. (2002). A new optical method for the noninvasive detection of minimal tissue alterations. *Phys Med Biol, 47*, 2095–2108.

Charvet, I., et al. (2004). Diagnosis and grading of gastritis by non-invasive optical analysis. *European Journal of Gastroenterology and Hepatology, 16*(11), 1189–1198.

Christensen, C., Anthony, S.D., & Roth, E.A. (2004). Seeing what's next. HBS Press.

Cuschieri, A. (2005). Laparoscopic surgery: Current status, issues and future developments. *Surgeon, 3*, 125–130, 132–133, 135–138.

De Lorijn, F., et al. (2005). Diagnosis of Hirschsprung's disease: A prospective, comparative accuracy study of common tests. *J Pediatr, 146*, 787–792.

Firilas, A.M., Jackson, R.J., & Smith, S.D. (1998). Minimally invasive surgery: The pediatric surgery experience. *J Am Coll Surg, 186*(5), 542–544.

Georgeson, K. (2003). Minimally invasive surgery in neonates. *Semin Neonatol, 8*, 243–248.

Georgeson, K.E., & Robertson, D.J. (2004). Laparoscopic-assisted approaches for the definitive surgery for Hirschsprung's disease. *Semin Pediatr Surg, 13*, 256–262.

Georgeson, K.E., & Robertson, D.J. (2004). Minimally invasive surgery in the neonate: Review of current evidence. *Semin Perinatol, 28*, 212–220.

Georgeson, KE., et al. (1999). Primary laparoscopic-assisted endorectal colon pull-through for Hirschsprung's disease: A new gold standard. *Annals of Surgery, Ann Surg, 229*, 678–683.

Gilger, M.A. (2001). Gastroenterologic endoscopy in children: Past, present, and future. *Curr Opin Pediatr, 13*: 429–434.

Given Imaging Web site. http://www.givenimaging.com

Goodman, C.S. (2004). Hta 101: Introduction to health technology assessment. Falls Church, VA: The Lewin Group.

Guilhon de Araujo Sant'Anna, A.M., Dubois, J., Miron, M.C., & Seidman, E.G. (2005). Wireless capsule endoscopy for obscure small-bowel disorders: Final results of the first pediatric controlled trial. *Clin Gastroenterol Hepatol, 3*, 264–270.

Hanly, E.J., & Talamini, M.A. (2004). Robotic abdominal surgery. *Am J Surg, 188*(Suppl 4A), 19S–26S.

Kant, A.J., Klein, M.D., & Langenburg, S.E. (2004). Robotics in pediatric surgery: Perspectives for imaging. *Pediatr Radiol, 34*, 454–461.

Kavin, H., Berman, J., Martin, T.L., Feldman, A., & Forsey-Koukol, K. (2006). Successful wireless capsule endoscopy for a 2.5-year-old child: Obscure gastrointestinal bleeding from mixed, juvenile, capillary hemangioma-angiomatosis of the jejunum. *Pediatrics, 117*, 539–543.

Martucciello, G., et al. (2005). Controversies concerning diagnostic guidelines for anomalies of the enteric nervous system: A report from the fourth International Symposium on Hirschsprung's disease and related neurocristopathies. *J Pediatr Surg, 40,* 1527–1531.

Mazziotti, M.V., & Langer, J.C. (2001). Laparoscopic full-thickness intestinal biopsies in children. *J Pediatr Gastroenterol Nutr, 33*, 54–57.

Menciassi, A., Moglia, A., Gorini, S., Pernorio, G., Stefanini, C., & Dario, P. (2005). Shape memory alloy clamping devices of a capsule for monitoring tasks in the gastrointestinal tract. *J Micromech Microeng, 15*, 2045–2055.

Moglia, A., Menciassi, A., Schurr, M.O., & Dario, P. (2006). Wireless capsule endoscopy: From diagnostic devices to multipurpose robotic systems. *Biomed Microdevices*, [published ahead of print].

Nagumo, J., et al. (1962). Echo capsule for medical use (a batteryless endoradiosonde). *IRE Trans Bio-Med Electronics, 9*, 195.

Nebeker, F. (2002). Golden accomplishments in biomedical engineering. *IEEE Eng Med Biol, 21*, 17–47.

Noblett, H.R. (1969). A rectal suction biopsy tube for use in the diagnosis of Hirschsprung's disease. *J Pediatr Surg, 4*, 406–409.

North American Society for Pediatric Gastroenterology Hepatology, and Nutrition Web site. http://www.naspghan.org/assets/diseaseInfo/pdf/Hirschsprung-E.pdf

Olives, J.P., et al., and the European Society for Paediatric Gastroenterology, Hepatology and Nutrition. (2004). Advances in endoscopy and other diagnostic techniques: Working group report of the second World Congress of Pediatric Gastroenterology, Hepatology, and Nutrition. *J Pediatr Gastroenterol Nutr, 39*(Suppl 2), S589–S595.

Prato, A.P., Martucciello, G., & Jasonni, V. (2001). Solo-RBT: A new instrument for rectal suction biopsies in the diagnosis of Hirschsprung's disease. *J Pediatr Surg, 36*, 1364–1366.

Rehbein, F., Morger, R., Kundert, S., & Meher-Ruge, W. (1966). Surgical problems in congenital mergacolon: A comparison of surgical techniques. *Journal of Pediatric Surgery, 1*(6), 526–532.

Ruckauer, K.D., & von Dobschuetz, E. (2005). Laparoscopically assisted colorectal resection in Hirschsprung's disease. *Eur J Med Res, 10*, 361–365.

San Luis, S., Nemoto, T., & Beardmore, H. (1968). Surgical treatment of Hirschsprung's disease. *Surgery, 63*(2), 331–338.

Seidman, E.G., Sant'Anna, A.M., & Dirks, M.H. (2004). Potential applications of wireless capsule endoscopy in the pediatric age group. *Gastrointest Endosc Clin N Am*, 14, 207–217.

Silbermintz, A., Sahdev, I., Moy, L., Vlachos, A., Lipton, J., & Levine, J. (2006). Capsule endoscopy as a diagnostic tool in the evaluation of graft-vs.-host disease. *Pediatr Transplant, 10*, 252–254.

Soaue, F. (1964). *Hirschsprungs disease: A new surgical approach*. Proceedings of the British Association of Paediatric Surgeons in Sheffield Surgeons, Sheffield.

Soper, R., & Figueroa, P. (1971). Surgical treatment of Hirschsprungs disease: Comparison of modifications of the Duhamel and Soaue operations. *Journal of Pediatric Surgery, 6*(6).

Woo, R., Le, D., Krummel, T.M., & Albanese, C. (2004). Robot-assisted pediatric surgery. *Am J Surg, 188*(Suppl 4A), 27S–37S.

Yagmurlu, A., Harmon, C.M., & Georgeson, K.E. (2006). Laparoscopic cecostomy button placement for the management of fecal incontinence in children with Hirschsprung's disease and anorectal anomalies. *Surg Endosc, 20*, 624–627.

Zitsman, J.L. (2003a). Current concepts in minimal access surgery for children. *Pediatrics, 111*, 1239–1252.

Zitsman, J.L. (2003b). Current concept in minimally access surgery for children. *Pediatrics, 111*(6).

Zitsman, J.L. (2006). Pediatric minimal-access surgery: Update 2006. *Pediatrics, 118*, 304–308.

KEY TERMS

Endoluminal Capsule Monitoring: A noninvasive medical diagnostic procedure that uses a miniaturized wireless radio transmitter embedded into an ingestible water-tight capsule to monitor the gastrointestinal system.

Endoscopy: Looking inside the human body for medical reasons.

Hirschsprung's Disease: Also known as congenital megacolon, occurs when some of the nerve cells in the intestines do not form properly.

Minimally Invasive Surgery (MIS): Surgery performed by entering the body through a small orifice.

Minimally Invasive Techniques (MIT): Techniques required to ensure that MIS can occur.

A Mobile Computing Framework for Passive RFID Detection System in Health Care

Masoud Mohammadian
University of Canberra, Australia

Ric Jentzsch
Compucat Research Pty Limited, Australia

INTRODUCTION

The cost of health care continues to be a world wide issue. Research continues into ways and how the utilization of evolving technologies can be applied to reduce costs and improve patient care, while maintaining patient's lives. To achieve these needs requires accurate, near real time data acquisition and analysis. At the same time there exists a need to acquire a profile on a patient and update that profile as fast and as possible. All types of confidentiality need to be addressed no matter which technology and application is used. One possible way to achieve this is to use a passive detection system that employs wireless radio frequency identification (RFID) technology. This detection system can integrate wireless networks for fast data acquisition and transmission, while maintaining the privacy issue. Once this data is obtained, then up to date profiling can be integrated into the patient care system. This article discussed the use and need for a passive RFID system for patient data acquisition in health care facilities such as a hospital. The development of profile data is assisted by a profiling intelligent software agent that is responsible for processing the raw data obtained through RFID and database and invoking the creation and update of the patient profile.

BACKGROUND

Health is on everyone's agenda whether they are old or young. Millions of hours of lost time is recorded each week by employers' whose staff are in need of health care. It is and has been known that more research into applications and innovative architectures is needed. To this end the use of Radio Frequency Identification (RFID), a relatively new technology and is showing itself to be a viable and promising technology as an aid to health care (Finkenzeller, 1999; Glover & Bhatt, 2006; Hedgepeth, 2007; Lahiri, 2005; Schuster, Allen, & Brock, 2007; Shepard, 2005). This technology has the capability to penetrate and add value to nearly every area of health care. It can be used to lower the cost of some services as well as improving service to individuals and the health care provider. Although many organizations are developing and testing the possible use of RFIDs, the real value of RFID is achieved in conjunction with the use of intelligent software agents. Thus the issue becomes the integration of these two great technologies for the benefit of assisting health care services.

To begin with, let us look at data collection. In health care, we can collect data on the patients, doctors, nurses, institution itself, drugs and prescriptions, diagnosis, and many other areas. It would not be feasible to do all of these nor would all of these be able to effectively use RFID. Thus for our perspective we will concentrate on a subset with the understanding that all areas could, directly or indirectly, benefit from the use of RFID and intelligent software agents in a health care and hospital environment.

In this research, we begin to look at the architecture of integrating intelligent software agents technology with RFID technology, in particular in managing patients' health care data in a hospital environment.

An intelligent software agent can continuously profile a patient based on their medical history, current illness, and on going diagnostics. The RFID provides the passive vehicle to obtain the data via its monitoring capabilities. The intelligent software agent provides the active vehicle in the interpretation profiling of the data and reporting capacity. There are certain data that is stored about each patient in a hospital. The investigation of this data provides an analysis that describes the patient's condition, is able to monitor their status, and cross reflect on why the patient was admitted to the hospital. Using this information an evolving profile of each patient can be constructed and analysed.

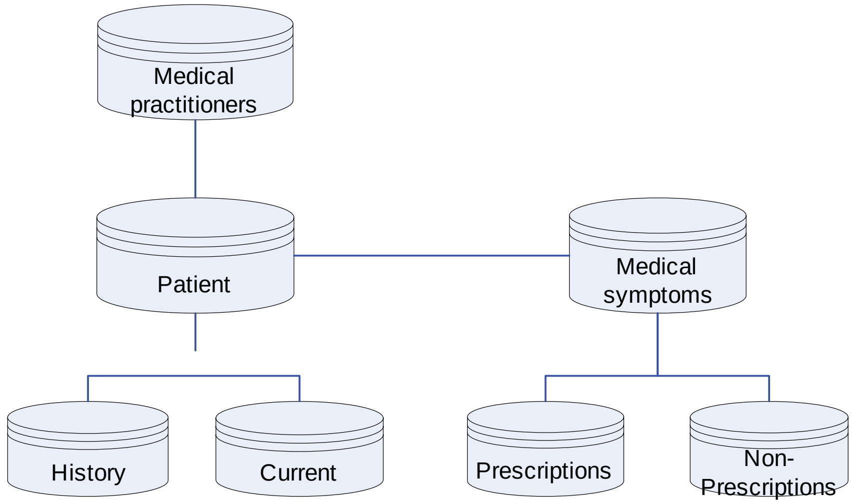
Using the data and analysis this will allow us to assist in deciding what kind of care he/she requires, the effects of ongoing care, and how to best care for this patient using available resources (doctors, nurses, beds, etc.) for the patient. The software agent is used to build a profile of each patient as they are admitted to the health care institution. Although not shown in the illustration, an additional profile for each doctor can be developed that practices in the hospital can be developed. If this is done, then the patient and doctor profile can be correlated to obtained the availability of the best doctor to suit the patient. However, this will require an additional data repository, as shown in Figure 1.

The patient profiling is useful in a variety of situations:

- The profile provides a personalized service based on the patient and not on symptoms or illness. to a particular patient. For example, by identifying the services that the patient requires this will allow us to target that which will be directed to speeding up their recovery progress;
- A good profile will assist the medical facilities in trying to prevent the need for the patient to return to the hospital any sooner than necessary;
- Disambiguating patient's diagnostic based on patient profile may help in assisting in matching a doctor's specialization to the right patient;
- When a patient needs to re-enter a hospital, a past profile can make it easier to match the patient's needs to a relevant available doctor;
- Presenting information about the patient on an on-going, continuous basis for the doctors means that current up to date information is available rather than information that needs to be searched for and compiled before it is useful; and
- Providing tailored and appropriate care to reduce health care costs.

Profiling is being done in many business operations today. Often profiling is combined with personalization, and user modeling for many e-commerce applications such as those by IBM, ATG Dynamo,

Figure 1. Data repositories for patient and doctors



BroadVision, Amazon, and Garden (18). However, there is very little in the way of the use of such systems in hospital and very little in health care in this perspective has been reported. It must be remember that there are different definitions of personalization, user modeling, and profiling. In e-commerce the practice of tracking information about consumers' interests by monitoring their movements online is considered profiling or user modeling. This can be done without using any personal information, but simply by analyzing the content, URL's, and other information about a user's browsing path/click-stream. Many user models try to predict the user's preference in a narrow and specific domain. This works well as long as that domain remains relatively static and, as such, the results of such work may be limited.

In this research, profiling is a technique whereby a set of characteristics of a particular class of person, patient, is inferred from their past and data-holdings are then searched for individuals with a close fit to symptom characteristics. One of the main aims of profiling and user modeling is to provide information recipients with correct and timely response for their needs. This entails an evolving profile to ensure that as the dynamics of that which is being monitored change, the profile and model reflects these updates as appropriate.

There are several ways in which a patient's visit to a hospital can be recorded. A patient's visit may simply be classified as a regular visit. This may be for a check up, for tests, or at the request of a doctor. A patient might be at the hospital because of an emergency or an ad hoc appointment due to lack of other facilities being available. Of course there are a whole set of patients that visit the hospital for reasons that are less well defined. In each situation, the needs of the patients are different.

The patient's profile can assist the attending doctor in being aware of the particular patient's situation. This provides the attending doctor with information that is needed without waiting for the patient's regular doctor. The regular doctor may be unavailable and therefore the profile of the patient can be matched with the available doctor suitable to the needs of the patient. The patient to doctor assignment is a type of scheduling issue and is not going to be discussed in this article.

However, in an emergency visit, there is no assigned doctor for such a patient. The doctor in emergency section of the hospital will provide information about a patient after examination and a patient profile then can be created. In this case, the intelligent agent can assist the patient by matching the profile of the patient with the doctors suitable to the needs of the patient. Also the doctors can be contacted in a speedier manner as they are identified and their availability is known.

An appointment visit is very similar to a regular visit but it may happen only once and therefore the advantages mentioned for regular patients applies here.

We will endeavor to describe several of these, but will expand on one particular potential use of RFIDs in managing patient health data. First let us provide some background on RFIDs and present some definitions. We will discuss the environment that RFIDs operate in and their relationship to other available wireless technologies such as the IEEE 802.11b, IEEE 802.11g, IEEE 802.11n, and so forth, in order to fulfill their requirements effectively and efficiently.

This research is divided into four main sections. Section two is based on the patient to doctor profiling and intelligent software agents. The third section is a RFID background; this will provide a good description of RFIDs and their components. This section discusses several practical cases of RFID technology in and around hospitals. It will also list three possible applicable cases assisting in managing patients' medical data. The final section discusses the important issue of maintaining patients' data security and integrity and relates that to RFIDs.

PATIENT TO DOCTOR PROFILING

A profile represents the extent to which something exhibits various characteristics. These characteristics are used to develop a linear model based on the consensus of multiple sets of data, generally over some period of time. A patient or doctor profile is a collection of information about a person based on the characteristics of that person. This information can be used in a decision analyze situation between the doctor, domain environment, and patient. The model can be used to provide meaningful information for useful and strategic actions. The profile can be static or dynamic. The static profile is kept in prefixed data fields where the period between data field updates is long such as months or years. The dynamic profile is constantly updated as per evaluation of the situation. The updates may be performed manually or automated. The automated user profile building is especially important in real time decision-making systems. Real time systems are dynamic. These systems often contain data that is critical to the user's decision making process. Manually updated profiles are at the need and discretion of the relevant decision maker.

The profiling of patient doctor model is base on the patient/doctor information. These are:

- The categories and subcategories of doctor specialization and categorization. These categories will assist in information processing and patient/doctor matching.
- Part of the patients profile based on symptoms (past history problems, dietary restrictions, etc.) can assist in prediction of the patient's needs specifically.
- The patients profile can be matched with the available doctor profiles to provide doctors with information about the arrival of patients as well as presentation of the patients profile to a suitable, available doctor.

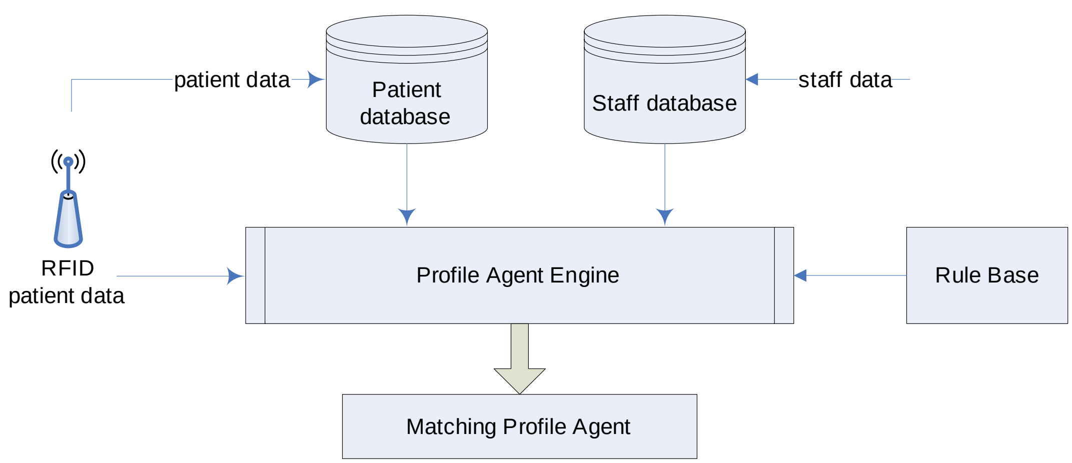
A value denoting the degree of association can be created form the above evaluation of the doctor to patient's profile. The intelligent agent based on the denoting degrees and appropriate, available doctors can be identified and allocated to the patient.

In the patient/doctor profiling, the agent will make distinctions in attribute values of the profiles and match the profiles with highest value. It should be noted that the agent creates the patient and doctor profiles based on data obtained from the doctors and patient namely:

- Explicit profiling occurs based on the data entered by hospital staff about a patient.
- Implicit profiling can fill that gap for the missing data by acquiring knowledge about the patient from its past visit or other relevant databases if any and then combining all these data to fill the missing data. Using legacy data for complementing and updating the user profile seems to be a better choice than implicit profiling. This approach capitalizes on user's personal history (previous data from previous visit to doctor or hospital).

The proposed agent architecture allows user profiling and matching in such a time intensive important application. The architecture of the agent profiling systems using RFID is given in Figure 2.

Figure 2. Agent profiling model using RFID



PROFILE MATCHING

Profile matching done is based on a vector of weighted attributes. To get this vector, a rule based systems can be used to match the patient's attributes (stored in patient's profile) against doctor's attributes (stored in doctor's profile). If there is a partial or full match between them, then the doctor will be informed (based on their availability from the hospital doctor database).

INTEGRATION OF INTELLIGENT SOFTWARE AGENTS AND RFID TECHNOLOGIES

Intelligent software agent technology has been used in order to provide the needed transformation of RIFD passive data collection into an active organizational knowledge assistant (Finkenzeller, 1999; Glover & Bhatt, 2006; Hedgepeth, 2007; Lahiri, 2005; Schuster et al., 2007; Shepard, 2005). Intelligent agent should be able to act on new data and already stored profile/knowledge and thereafter to examine its current actions based on certain assumptions, and inferentially plans its activities. Furthermore, intelligent software agents must be able to *interact* with other agents using symbolic language (Bigus & Bigus, 1998; Wooldridge & Jennings, 1995) and able to substitute for a range of human activities in a situated context. (In our case the activities are medical/patient assignment and the context is a hospital environment)

Context driven Intelligent software agents' activities are also dynamic and under continuous development in an historical time related environment (Bigus & Bigus, 1998; Wooldridge & Jennings, 1995).

Medical and hospital patient applied ontology's describing the applied domain are necessary for the semantic communication and data understanding between RFID inputs and knowledge bases inference engines so that profiling of both patients and doctors can be achieved (Gruber, 1993; Guarino, Carrara, & Giaretta, 1994).

The integration of RFID capabilities and intelligent agent techniques provides promising development in the areas of performance improvements in RFID data collection, inference, knowledge acquisition, and profiling operations.

By using mediated activity theories, an RIFD agent architecture could be modeled according to the following characteristics:

- The ability to use patient/doctor profile in natural language, ACL, or symbolic form as communicative tools mediating agents cooperative activities.
- The ability to use subjective and objective properties required by intelligent software agents to perform bidirectional multiple communication activities.
- The ability to internalize representations of medical/patient profile patterns from agents or humans.
- The ability to externalize internally stored representations of medical assignment patterns to other agents or humans.

The Agent Language Mediated Activity Model (ALMA) agent architecture currently under research is based on the mediated activity framework described and is able to provide RFID with the necessary framework to profile a range of internal and external medical/patient profiling communication activities performed by wireless multi-agents.

RFID DESCRIPTION

RFID or Radio Frequency Identification is a progressive technology that has been said to be easy to use and well suited for collaboration with intelligent software agents. Basically an RFID can:

- Be read-only;
- Volatile read/write; or
- Write once/read many times
- RFID are:
  - Noncontact and
  - Non line-of-sight operations.

Being noncontact and non line-of-sight will make RFIDs able to function under a variety of environmental conditions and while still providing a high level of data integrity (Finkenzeller, 1999; Glover & Bhatt, 2006; Hedgepeth, 2007; Lahiri, 2005; Schuster et al., 2007; Shepard, 2005).

MAIN COMPONENTS

A basic RFID system consists of four components:

1. The RFID tag (sometimes referred to as the transponder);
2. A coiled antenna;
3. A radio frequency transceiver; and
4. Some type of reader for the data collection.

Basically there are three components as often components are combined such as the transponder or transceiver or the antenna.

Transponders

The reader emits radio waves in ranges of anywhere from 2.54 centimeters to 33 meters. Depending upon the reader's power output and the radio frequency used and if a booster is added that distance can be somewhat increased. When RFID tags (transponders) pass through a specifically created electromagnetic zone, they detect the reader's activation signal. Transponders can be online or off-line and electronically programmed with unique information for a specific application or purpose. A reader decodes the data encoded on the tag's integrated circuit and passes the data to a server for data storage or further processing.

Coiled Antenna

The coiled antenna is used to emit radio signals to activate the tag and read or write data to it. Antennas are the conduits between the tag and the transceiver that controls the system's data acquisition and communication. RFID antennas are available in many shapes and sizes. They can be built into a doorframe, book binding, DVD case, mounted on a tollbooth, embedded into a manufactured item such as a shaver or software case (just about anything) so that the receiver tags the data from things passing through its zone (Finkenzeller, 1999; Glover & Bhatt, 2006; Hedgepeth, 2007; Lahiri, 2005; Schuster et al., 2007; Shepard, 2005).

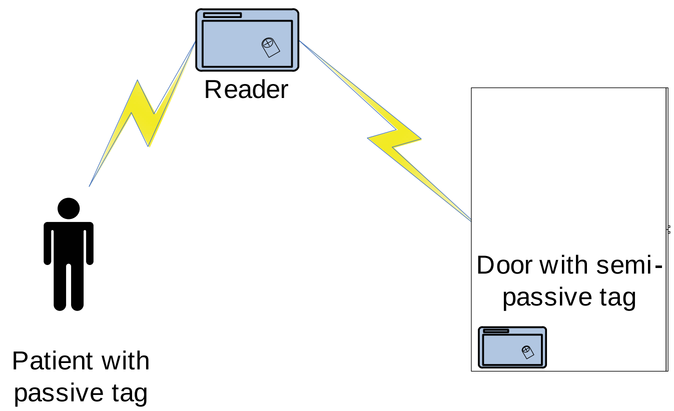
Transceiver

Often the antenna is packaged with the transceiver and decoder to become a reader. The decoder device can be configured either as a handheld or a fixed-mounted device. In large complex, often chaotic environments, portable or handheld transceivers would prove valuable.

TYPES OF RFID TRANSPONDERS

RFID tags can be categorized as active, semi-active, or passive. Each has and is being used in a variety of inventory management and data collection applications today. The condition of the application, place and use determines the required tag type.

Figure 3. Semi-passive tag



Active Tags

Active RFID tags are powered by an internal battery and are typically read / write. Tag data can be rewritten and/or modified as the need dictates. An active tag's memory size varies according to manufacturing specifications and application requirements; some tags operate with up to 5 megabyte of memory. For a typical read/write RFID work-in-process system, a tag might give a machine a set of instructions, and the machine would then report its performance to the tag. This encoded data would then become part of the tagged part's history. The battery-supplied power of an active tag generally gives it a longer read range. The trade off is greater size, greater cost, and a limited operational life that has been estimated to be a maximum of 10 years, depending upon operating temperatures and battery type (Finkenzeller, 1999; Glover & Bhatt, 2006).

Semi-Active Tags

The semi-active tag comes with a battery. The battery is used to power the tags circuitry and not to communicate with the reader. This makes the semi-active tag more independent than the passive tag, and it can operate in more adverse conditions. The semi-active tag also has a longer range and more capabilities than a passive tag (Shepard, 2005). Linear barcodes that reference a database to get product specifications and pricing are also data devices that act is a very similar way. Semi-passive tags are preprogrammed, but can allow for slight modifications of their instructions via the reader/interrogator. However, it is bigger, weighs more, and is more complete than a passive tag. A reader is still needed for data collection.

Passive Tags

Passive RFID tags operate without a separate external power source and obtain operating power generated from the reader. Passive tags, since they have no power source embedded, are consequently much lighter than active tags, less expensive, and offer a virtually unlimited operational lifetime. However, the trade off is that they have shorter read ranges, than active tags, and require a higher-powered reader.

Read-only tags are typically passive and are programmed with a unique set of data (usually 32 to 128 bits) that cannot be modified. Read-only tags most often operate as a data device that utilizes a database for all data storage (Finkenzeller, 1999; Shepard, 2005).

Range

RFID systems can be distinguished by their deployment and frequency range. RFID tags generally operate in two different types of frequencies that make them adaptable for nearly any application. These frequency ranges are:

Low Frequency Range (Short Range)

Low-frequency (30 KHz to 500 KHz) systems have short reading ranges and lower system costs. They are most commonly used in security access, asset tracking, and animal identification applications (Glover & Bhatt, 2006; Hedgepeth, 2007; Lahiri, 2005).

High Frequency Range (Long Range)

High-frequency (850 MHz to 950 MHz and Industry, Science and Medical - 2.4 GHz to 2.5 GHz) systems, offer longer reading ranges (greater than 33 meters) and high reading speeds. These systems are generally used for such applications as railroad car tracking, container dock and transport management, and automated toll collection. However, the higher performance of high-frequency RFID systems incurs higher system operating costs (Glover & Bhatt, 2006; Schuster et al., 2007).

Hospital Environment

In hospitals, systems need to use rules and domain knowledge that is appropriate to the situation. One of the more promising capabilities of intelligent software agents is their ability to coordinate information between the various resources.

In a hospital environment, in order to manage patient medical data we need both types; fixed and handheld transceivers. Also, transceivers can be assembled in ceilings, walls, or doorframes to collect and disseminate data. Hospitals have become large complex environments.

In a hospital, nurses and physicians can retrieve the patient's medical data stored in transponders (RFID tags) before they stand beside a patient's bed or as they are entering a ward.

Given the descriptions of the two types and their potential use in hospital patient data management we suggest that:

- It would be most useful to embed a passive RFID transponder into a patient's hospital wrist band;
- It would be most useful to embed a passive RFID transponder into a patient's medical file (there are several versions and perspectives that we can take no this).

Doctors should have PDAs equipped with RFID or some type of personal area network device. Either would enable them to retrieve some patient's information whenever they are near the patient, instead of waiting until the medical data is pushed to them through the hospital server (there are several versions and perspectives that we can take no this):

- *Active RFID tags* are more appropriate for the continuous collection of the patient's medical data. Since the patient's medical data needs to be continuously recorded to an active RFID tag and an associated reader needs to be employed. Using an active RFID means that the tag will be a bit bulky because of the needed battery for the write process and there is a concern with radio frequency admissions. Thus, it is felt that an active tag would not be a good candidate for the patient wrist band. However, if the patient's condition is to be continuously monitored, the collection of the data at the source is essential. The inclusion of the tag in the wrist band is the only way to recorder the medical data on a real-time base using the RFID technology. As more organizations get into the business of manufacturing RFIDs and the life and size of batteries decrease, the tag size will decrease and this may be a real possible use.
- *Passive RFID tags* can be also used as well. These passive tags can be embedded in the doctors PDA, which is needed for determining their locations whenever the medical staff requires them. Also, passive tags can be used in patients' wrist bands for storage of limited amount of data- on off-line bases, for example, date of hospital admission, medical record number, and so forth.

After examining both ranges, we can suggest the following:

- *Low frequency range tags* are suitable for the patients' band wrist RFID tags. Since we expect that the patients' bed will not be too far from a RFID reader. The reader might be fixed over the patient's bed, in the bed itself, or over the door-frame. The doctor using his/her PDA would be aiming to read the patient's data directly and within a relatively short distance.
- *High frequency range tags* are suitable for the physician's tag implanted in their PDAs. As physicians use to move from one location to another in the hospital, data on their patients could be continuously being updated.

One final point in regards to the range of RFIDs: until 2002, the permissible radio frequency range was not regulated, that is, it still operated in some low frequency ranges (30- 500 KHz) and in the free 2.45 GHz ISM band of frequency. The IEEE's 802.11b and IEEE's 802.11g (WiFi) wireless networks also operate in the same range (actually there are many other wireless application that operate in that range). This band of frequency is crowded. Where equipment in a hospital is often in the ISM band of frequency, there may be some speed of transmission degradation. The IEEE 802.11n builds upon previous 802.11 standards by adding MIMO (multiple-input multiple-output). MIMO uses multiple transmitter and receiver antennas to allow for increased data throughput via spatial multiplexing and increased range by exploiting the spatial diversity. Note that 802.11n draft 2.0 has been released but the certification of products is still in progress. What this means is that even though 802.11n has greater benefits then previous standards, it is still a draft. The full version is not expected until 2008; thus; products may take several years to be compliant and incorporate that into RFIDs (Hedgepeth, 2007).

Shapes of RFID Tags

RFID tags come in a wide variety of shapes and sizes. Animal tracking tags that are inserted beneath the animal's skin can be as small as a pencil lead in diameter and about one centimeter in length. Tags can be screw-shaped to identify trees or wooden items, or credit card shaped for use in access applications. The antitheft hard plastic tags attached to merchandise in stores are RFID tags (Glover & Bhatt, 2006). Manufacturers can create the shape that is best for the application, including flexible shaped tags that act like and resemble human skin. RFID tags can be flexible and do not have to be rigid.

Transceivers

The transceivers/interrogators can differ quite considerably in complexity, depending upon the type of tags being supported and the application. T\he overall function of the application is to provide the means of communicating with the tags and facilitating data transfer. Functions performed by the reader may include quite sophisticated signal conditioning, parity error checking, and correction. Once the signal from a transponder has been correctly received and decoded, algorithms may be applied to decide whether the signal is a repeat transmission, and may then instruct the transponder to cease transmitting or temporarily cease asking for data from the transponder. This is known as the "Command Response Protocol" and is used to circumvent the problem of reading multiple tags over a short time frame. Using interrogators in this way is sometimes referred to as "Hands Down Polling." An alternative, more secure, but slower tag polling technique is called "Hands Up Polling." This involves the transceiver looking for tags with specific identities, and interrogating them in turn. A further approach may use multiple transceivers, multiplexed into one interrogator, but with attendant increases in costs (Glover & Bhatt, 2006; Hedgepeth, 2007; Lahiri, 2005; Schuster et al., 2007).

Hospital patient data management deals with sensitive and critical information (patient's medical data). *Hands Down polling* techniques in conjunction with multiple transceivers that are multiplexed with each other, form a wireless network. The reason behind this choice is that, we need high speed for transferring medical data from medical equipment to or from the RFID wrist band tag to the nearest RFID reader, and then through a wireless network or a network of RFID transceivers or LANs to the hospital server. From there it is a short distance to be transmitted to the doctor's PDA, a laptop, or desktop through a WLAN IEEE 802.11b, 802.11g, or 802.11n, or wired LAN which operates at the 5.2 GHz band with a maximum data transfer rate exceeding 104 Mbps.

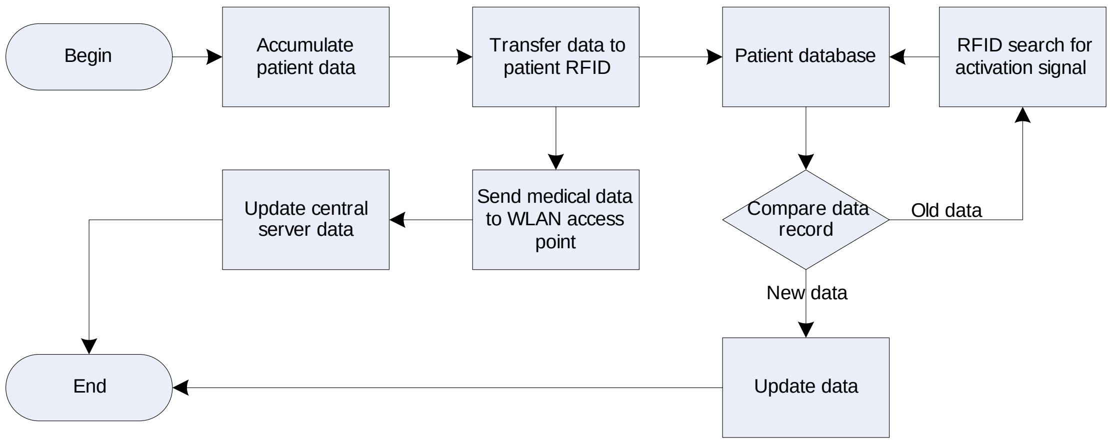
The "Hand Down Polling" techniques, as previously described, provides the ability to detect all detectable RFID tags at once (i.e., in parallel). Preventing any unwanted delay in transmitting medical data corresponding to each RF tagged patient.

RFID TRANSPONDER PROGRAMMERS

Transponder programmers are the means, by which data is delivered to write once, read many (WORM) and read/write tags. Programming can be carried out off-line or online. For some systems re-programming may be carried out online, particularly if it is being used as an interactive portable data file within a production environment, for example. Data may need to be recorded during each process. Removing the transponder at the end of each process to read the previous process data, and to program the new data, would naturally increase process time and would detract substantially from the intended flexibility of the application. By combining the functions of a transceiver and a programmer, data may be appended or altered in the transponder as required, without compromising the production line.

We conclude from this section that RFID systems differ in type, shape, and range; depending on the type of application, the RFID components shall be chosen. Low frequency range tags are suitable for the patients' band wrist RFID tags. Since we expect that a patients' bed is not too far from the RFID reader, which might be fixed on the room ceiling or door-frame. High frequency range tags are suitable for the physician's PDA tag. As physicians use to move from on location to another in the hospital, long read ranges are required. On the other hand, transceivers which deal with sensitive and critical information (patient's medical data) need the Hands Down polling techniques.

Figure 4. Acquisition of patient data



These multiple transceivers should be multiplexed with each other forming a wireless network.

PRACTICAL CASES USING RFID TECHNOLOGY

This section explains in details three possible applications of the RFID technology in three applicable cases. Each case is discussed step-by-step then represented by a flowchart. Those cases cover issues as acquisition of Patient's Medical Data, locating the nearest available doctor to the patients location, and how doctors stimulate the patient's active RFID tag using their PDAs in order to acquire the medical data stored in it.

Case I: Acquisition of Patient's Medical Data

Case one will represent the method of acquisition and transmission of medical data. This process can be described in the following points as follows:

1. A biomedical device equipped with an embedded RFID transceiver and programmer will detect and measure the biological state of a patient. This medical data can be an ECG, EEG, BP, sugar level, temperature or any other biomedical reading.

   After the acquisition of the required medical data, the biomedical device will write-burn this data to the RFID transceiver's EEPROM using the built in RFID programmer. Then the RFID transceiver with its antenna will be used to transmit the stored medical data in the EEPROM to the EEPROM in the patient's transponder (tag) which is around his/her wrist. The data received will be updated periodically once new fresh readings are available by the biomedical device. Hence, the newly sent data by the RFID transceiver will be accumulated to the old data in the tag. The purpose of the data stored in the patient's tag is to make it easy for the doctor to obtain medical information regarding the patient directly via the doctor's PDA, tablet PC, or laptop.
2. Similarly, the biomedical device will also transfer the measured medical data wirelessly to the nearest WLAN access point. Since high data rate transfer rate is crucial in transferring medical data, IEEE 802.11b or g is recommended for the transmission purpose.
3. The wirelessly sent data will be routed to the hospitals main server; to be then sent (pushed) to:
   i. Other doctors available throughout the hospital so they can be notified of any newly received medical data.
   ii. To an online patient monitoring unit or a nurse's workstation within the hospital.

iii. Or the acquired patients' medical data can be fed into an expert (intelligent) software system running on the hospital server. To be then compared with other previously stored abnormal patterns of medical data, and to raise an alarm if any abnormality is discovered.

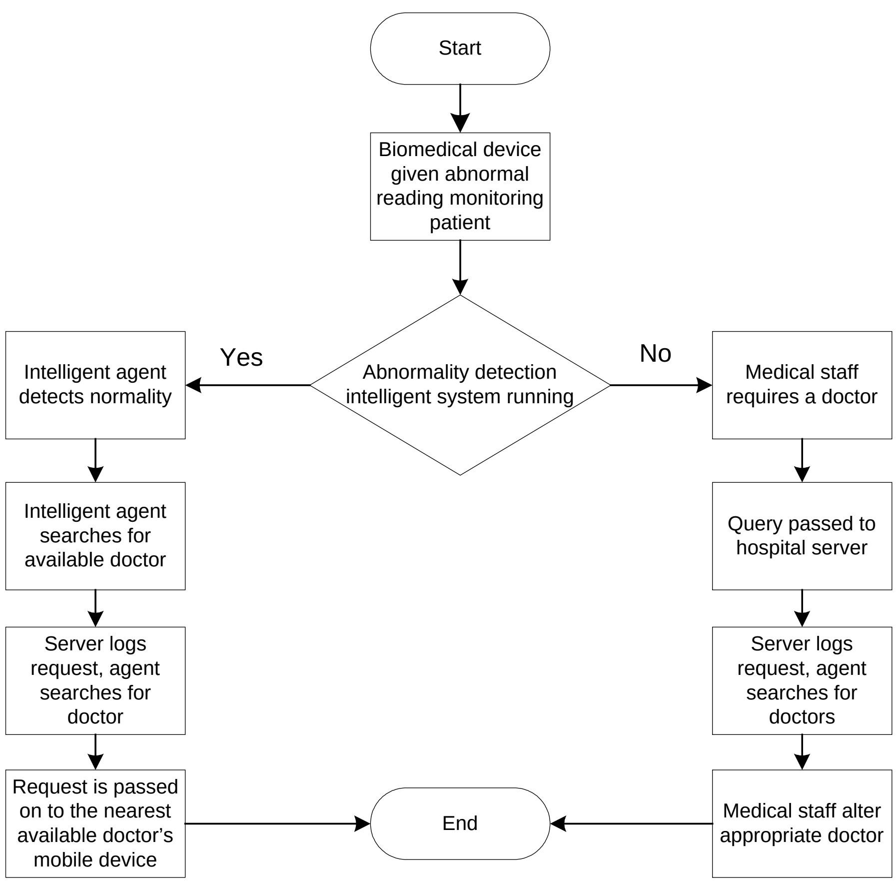
4. Another option could be using the in-built-embedded RFID transceiver in the biomedical device to send the acquired medical data wirelessly to the nearest RFID transceiver in the room. Then the data will travel simultaneously in a network of RFID transceivers until reaching the hospital server.

Case II: Locating the Nearest Available Doctor to the Patients Location

This case will explain how to locate the nearest doctor who is needed urgently to attend an emergency medical situation. This case can be explained as follows:

1. If a specific surgeon or physician is needed in a specific hospital department, the medical staff in the monitoring unit (e.g., nurses) can query the hospital server for the nearest available doctor to the patient's location. In our framework an intelligent agent can perform this task.
2. The hospital server traces all doctors' locations in the hospital through detecting the presences of their wireless mobile device; for example, PDA, tablet PC, or laptop in the WLAN range.
3. Another method that the hospital's server can use to locate the physicians is making use of the

Figure 5. Locating nearest doctor



RFID transceivers built-in the doctor's wireless mobile device. Similarly to the access points used in WLAN, RFID transceivers can assist in serving a similar role of locating doctor's location. This can be described in three steps, which are:

i. The fixed RFID transceivers throughout the hospital will send a stimulation signal to detect other free RFID transceivers—which are in the doctors PDAs, tablets , r laptops, and so forth.
ii. All free RFID transceivers will receive the stimulation signal and reply back with an acknowledgement signal to the nearest fixed RFID transceiver.
iii. Finally, each free RFID transceiver cell position would be determined by locating to which fixed RFID transceiver range it belongs to or currently operating in.

4. After the hospital server located positions of all available doctors, it determines the nearest requested physician (pediatrics, neurologist, and so forth) to the patient's location.
5. Once the required physician is located, an alert message will be sent to his\her PDA, tablet PC or laptop indicating the location to be reached immediately. This alert message could show:
   i. The building, floor, and room of the patient (e.g., 3C109).
   ii. Patient's case (e.g., heart stroke, arrhythmia, etc.)
   iii. A brief description of the patient's case.
6. If the hospital is running an intelligent agent as described in the proposed framework on its server, the process of locating and sending an alert message can be automated. This is done through comparing the collected medical data with previously stored abnormal patterns of medical data, then sending an automated message describing the situation. This system could be used instead of the staff in the patient monitoring unit or the nurse's workstation where nurses observe and then send an alert message manually.

Case III: Doctors Stimulate the Patient's Active RFID Tag Using their PDAs in Order to Acquire the Medical Data Stored in it

This method can be used in order to get rid of medical files and records placed in front of the patient's bed. Additionally, it could help in preventing medical errors- reading the wrong file for the wrong patient and could be considered as an important step towards a paperless hospital.

This case can be described in the following steps:

1. The doctor enters into the patient's room or ward. The doctor wants to check the medical status of a certain patient. So instead of picking up the "hard" paper medical file, the doctor interrogates the patient's RFID wrist tag with his RFID transceiver equipped in his\her PDA, tablet PC, or laptop, and so forth.
2. The patient's RFID wrist tag detects the signal of the doctor's RFID transceiver coming from his\her wireless mobile device and replies back with the patient's information and medical data.
3. If there was more than one patient in the ward possessing RFID wrist tags, all tags can respond in parallel using Hands Down polling techniques back to the doctor's wireless mobile device.
4. Another option could be that the doctor retrieving only the patient's number from the *passive* RFID wrist tag. Then through the WLAN the doctor could access the patient's medical record from the hospital's main server.

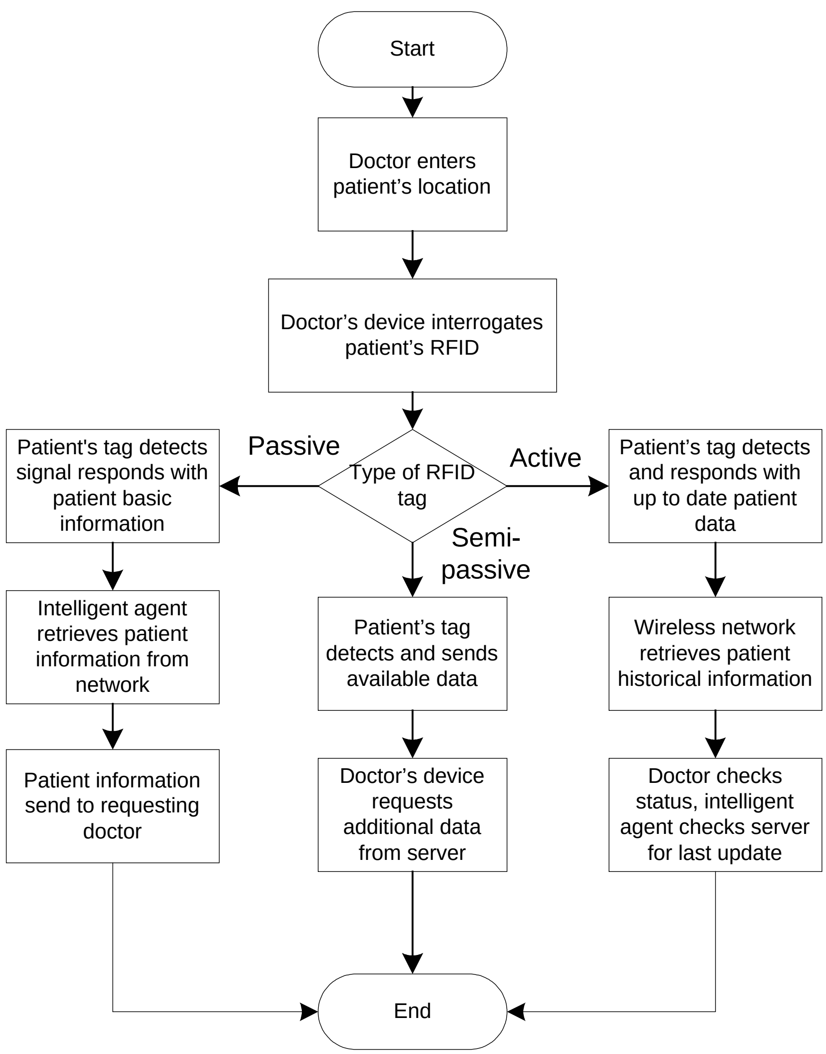
RFID technology has many potential important applications in hospitals, and the discussed three cases are a real practical example. Two important issues can be concluded from this section: WLAN is preferred for data transfer; given that IEEE's wireless networks have much faster speed and coverage area as compared to RFID transceivers\transponders technology. Yet, RFID technology is the best for data storage and locating positions of medical staff and patients as well.

The other point is that we need a RFID transceiver & programmer embedded in a biomedical device for data acquisition and dissemination, and only a RFID transceiver embedded in the doctor's wireless mobile device for obtaining the medical data. With the progress the RFID technology is currently gaining, it could become a standard as other wireless technologies (Bluetooth, for example), and eventually manufacturers building them in electronic devices; biomedical devices for our case.

MAINTAINING PATIENTS' DATA SECURITY AND INTEGRITY

Once data is transmitted wirelessly, security becomes a crucial issue. Unlike wired transmission, wirelessly transmitted data can be easily sniffed out leaving the transmitted data vulnerable to many types of attacks. For example, wireless data could be easily eavesdropped on using any mobile device equipped with a wireless card. In worst cases wirelessly transmitted data could be intercepted and then possibly tampered with, or in best cases, the patient's security and privacy would be compromised. Hence emerges the need for data to be initially encrypted from the source.

Figure 6. Functional flow



In this section, a discussion on how to apply encryption for the designed wireless framework for hospital is considered. Suggesting exactly where data needs to be encrypted and\or decrypted depending on the case that is being examined does this.

First a definition of the type of encryption that would be used in the design of the security (encryption\decryption) framework is discussed, followed by a flowchart demonstrating the framework in a step-by-step process.

Layers of Encryption

Two main layers of encryption are recommended. They are:

Physical (Hardware) Layer Encryption

This means encrypting all collected medical data at the source or hardware level before transmitting it. Thus, we insure that the patient's medical data would not be compromised once exposed to the outer world on its way to its destination. So even if a person with a malicious intent and also possessing a wireless mobile device steps into the coverage range of the hospitals' WLAN, this intruder will gain actually nothing since all medical data is encrypted, making all intercepted data worthless.

Application (Software) Layer Encryption

This means encrypting all collected medical data at the destination or application level once receiving it. Application level encryption runs on the doctor's wireless mobile device (e.g., PDA, tablet PC, or laptop) and on the hospital server. Once the medical data is received, it will be protected by a secret pass-phrase (encryption\decryption key) created by the doctor who possesses this device. This type of encryption would prevent any person from accessing patient's medical data if the doctor's wireless mobile device gets lost, or even if a hacker hacks into the hospital server via the Internet, intranet or some other mean.

Framework of Encrypting Patient's Medical Data

The previous section (Practical Cases using RFID Technology) focused on how to design a wireless framework to reflect how patients' medical data can be managed efficiently and effectively leading to the elimination of errors, delays, and even paperwork. Similarly, this section will focus on the previously discussed framework from a security perspective, attempting to increase security and data integrity.

i. Acquisition of Patients' Medical Data
ii. Doctors stimulating the patient's active RFID tag using their wireless mobile devices in order to acquire the medical data stored in it.

While the third case which was about locating the nearest available doctor to the patients location, is more concerned about locating doctors than transferring patient's data, so it is not discussed here.

The lower part of Figure 6 represents the physical (hardware) encryption layer. This part is divided into two sides. The left side demonstrates the case of a doctor acquiring patient's medical data via a passive RFID tag located in a band around the patient's wrist. The passive RFID tag contains only a very limited amount of information such as the patients name, date of admission to the hospital and above all his/her medical record number (MRN), which will grant access to the medical record containing the acquired medical data and other information regarding the patient's medical condition. This process is implemented in six steps, and involves two pairs of encryption and decryption. The first encryption occurs after the doctor stimulates the RFID passive tag to acquire the patient's MRN, so the tag will encrypt and reply back the MRN to the doctors PDA for example. Then the doctor will decrypt the MRN and use it to access the patient's medical record from the hospital's server. Finally the hospital server will encrypt and reply back the medical record, which will be decrypted once received by the doctors' PDA.

The right side of Figure 6 represents a similar case but this time using an active RF tag. This process involves only one encryption and decryption. The encryption happens after the doctor stimulates the active RFID tag using his PDA which has an in-equipped RFID transceiver, so the tag replies with the medical data encrypted. Then the received data is decrypted through the doctors' PDA.

The upper part of Figure 6 represents the application encryption layer, requiring the doctor to enter a pass-phrase to decrypt and then access the stored medical data. Whenever the doctor wants to access patient's medical data, the doctor simply enters a certain pass-phrase to grant access to either wireless mobile device or a hospital server depend where the medical data actually resides.

CHOOSING LEVEL OF SECURITY FOR THE WIRELESSLY-TRANSMITTED MEDICAL DATA

Securing medial data seems to be uncomplicated, yet the main danger of compromising such data comes from the people managing it, for example, doctors, nurses, and other medical staff. For that, we have seen that even though the transmitted medical data is initially encrypted from the source, doctors have to run application level encryption on their wireless mobile devices in order to protect this important data if the devices gets lost, left behind, robbed, and so forth. Nevertheless, there is a compromise. Increasing security through using multiple layers, and increasing length of encryption keys decreases the encryption\decryption speed and causes unwanted time delays, whether we were using application or hardware level of encryption. As a result, this could delay medical data sent to doctors or online monitoring units.

Figure 6 represents the case of high and low level of security in a flowchart applied to the previously discussed two cases in the last report.

At the end of this section, we conclude that there are two possible levels of encryption, software level (application layer) or hardware level (physical layer), depending on the level of security required. Both physical (hardware) layer and application layer encryption are needed in maintaining collected medical data on hospital servers and doctors wireless mobile devices.

Encrypting medical data makes the process of data transmission slower while sending data unencrypted is faster. We have to have a compromise between speed and security. For our case, medical data has to be sent as fast as possible to medical staff, yet the security issue has the priority.

CONCLUSION

Managing patients' data wirelessly (paperless) can prevent errors, enforce standards, make staff more efficient, simplify record keeping, and improve patient care. In this research report, both passive and active RFID tags were used in acquiring and storage of medical data, and then linked to the hospitals' server via a wireless network Moreover, three practical applicable RFID cases discussed how the RFID technology can be put in use in hospitals, while at the same time maintaining the acquired patients' data security and integrity.

This research in the wireless medical environment introduces some new ideas in conjunction to what is already available in RFID technology and wireless networks. Linking both technologies to achieve the research main goal, delivering patients medical data as fast and secure as possible, to pave the way for future paperless hospitals.

Finally, as reported by Frost and Sullivan, the high cost of radio frequency identification (RFID) technology is a deterrent for health care providers, though RFID has great benefits to hospitals in tracking patients, monitoring patients, assisting in health care administration, and reducing medical costs. With the reduction in cost of radio frequency identification (RFID) technology, increased use of RFID technology in health care in monitoring patients and assisting in health care administration is expected.

REFERENCES

Bigus, J. P., & Bigus, J. (1998). *Constructing intelligent software agents with Java – a programmers guide to smarter applications.* Wiley. ISBN: 0-471-19135-3.

Finkenzeller, K. (1999). *RFID handbook.* John Wiley and Sons Ltd.

Glover, B., & Bhatt, H. (2006). *RFID essentials.* O'Reilly Media, Inc. ISBN: 10-0596009445.

Gruber, T.R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition, 5,* 199-220.

Guarino, N., Carrara, M., & Giaretta, P. (1994). An ontology of meta-level categories. Journal of knowledge representation and reasoning. In *Proceedings of* *the Fourth International Conference (KR94)*, Morgan Kaufmann, San Mateo, CA.

Hedgepeth, W. O. (2007). *RFID metrics: Decision making tools for today's supply chains*. Boca Raton, FL: CRC Press. ISBN: 9780849379796.

Lahiri, S. (2005). *RFID sourcebook*. IBM Press. ISBN: 10-0131851373.

Odell, J. (Ed.). (2000, September). *Agent technology.* OMG Document 00-09-01, OMG Agents interest Group.

RFID Australia. (2003). *Why use RFID*. Retrieved February 15, 2008, from http://www.rfid-australia.com/files/htm/rfid%20brochure/page4.html

Schuster, E. W., Allen, S. J., & Brock D. L. (2007). *Global RFID: The value of the EPCglobal network for supply chain management*. Berlin; New York: Springer. ISBN: 9783540356547.

Shepard, S. (2005). *RFID: Radio frequency identification.* New York: McGraw-Hill. ISBN: 0071442995.

Wooldridge, M., & Jennings, N. (1995). Intelligent software agents: Theory and Practice. *The Knowledge Engineering Review*, 10(2), 115-152.

A Model for the Discussion of Medical Tourism

Mary Ann Keogh Hoss
Eastern Washington University, USA

Harm-Jan Steenhuis
Eastern Washington University, USA

INTRODUCTION

In recent years, the phenomenon of medical tourism has attracted increased media attention. Similar to manufacturing outsourcing where activities (and jobs) leave the United States, for example, because they can be conducted cheaper elsewhere, medical tourism deals with medical activities that can be conducted cheaper elsewhere. The phenomenon of medical tourism is relatively new, and scientific studies on this topic are rare. The objective in this chapter is to explore the issue of medical tourism. This chapter takes the viewpoint of the United States: What are the potential consequences of medical tourism for the United States?

BACKGROUND

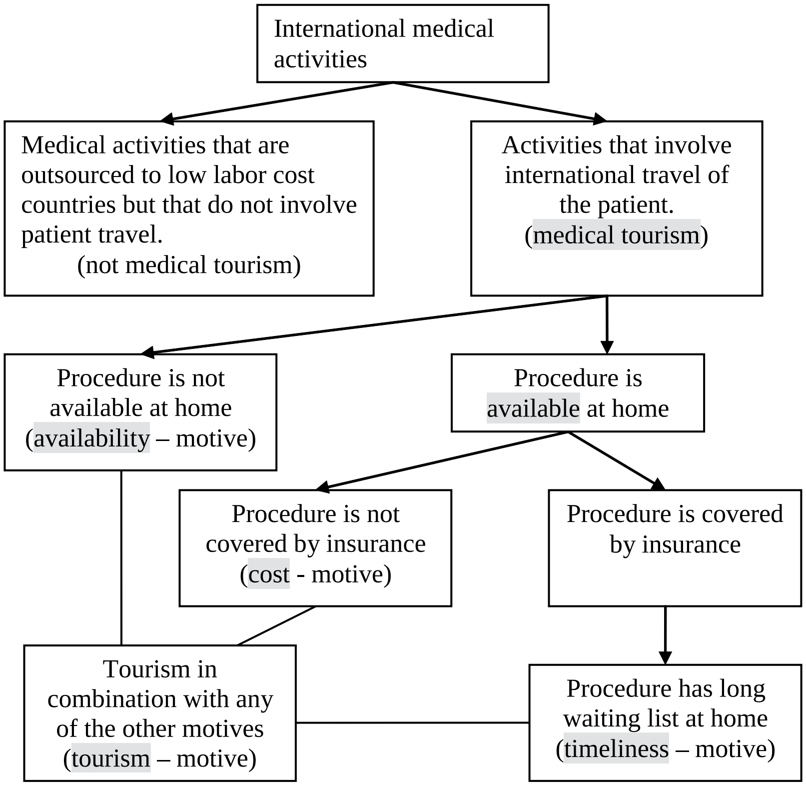
There are different types of medical tourism. In the past, medical tourism may have occurred primarily due to the lack of treatment or procedures at home; for instance, as a consequence of laws. Other types of motivation have occurred. One of these is the cost involved in the procedure. CBC News reported that many medical tourists from the United States are seeking treatment at a quarter or sometimes even a tenth of the cost at home (CBC News Online, 2004). Another motivation is waiting lines for treatments. For yet others, medical tourism may provide a chance to combine a tropical vacation with elective or plastic surgery. Figure 1 provides a theoretical framework for this analysis.

As Figure 1 illustrates, medical tourism is considered in *those situations where individuals travel abroad for medical treatment or procedures*. These individuals (the patients) may have several types of motives for traveling abroad. The direction for travel may depend on the motive. Offshore diagnostics is not included in the definition of medical tourism. Instances where medical analyses are outsourced to India because of lower costs but without the actual travel of the patients is not included as medical tourism. These instances may affect the U.S. health care industry. When medical tourists travel from the United States to a developing country like India, this doesn't mean that the care in India is of a lower quality. Doheny (2006) illustrates that if one chooses wisely, the care will be as good or even better than it is in the United States. Some overseas hospitals are accredited by the U.S.-based Joint Commission International (JCI), which is affiliated with the Joint Commission on Accreditation of Health care Organizations, which accredits U.S. hospitals (Doheny, 2006). Some countries such as Brazil, Cuba, Costa Rica, Hungary, India, Israel, Jordan, Lithuania, Malaysia, and Thailand are actively promoting medical tourism and are turning to this type of new industry in an attempt to generate hard currency (Bradley & Kim, 1994; Xinhua News Agency, 2005).

Availability Motive

The issue of availability has occurred with regard to abortions or artificial insemination (English, Mussell, Sheather & Sommerville, 2005). These topics were

Figure 1. Motives for medical tourism



related to the laws in specific countries that would not allow certain procedures to take place. Individuals could opt to go to other countries with different laws that would allow them to undergo a specific procedure. The travel may be primarily from less developed areas with either fewer treatments or procedures available due to lacking technology and/or local laws. It is unlikely at this time that this motive plays a major role for U.S. residents since most medical procedures are available in the United States.

Cost Motive

Another motive for medical tourism is the cost of the procedure or treatment. For example, some medical treatments may be cheaper in low labor cost countries. Doheny (2006) states that a U.S. surgeon's fee for a full face-lift averaged $4,822 in 2004. In the United States, this type of treatment does not include a hospital stay, and therefore, there are no hospital costs involved. At Bumrungrad Hospital in Bangkok, the surgeon's fee for a full face-lift is about $1,200, and this includes a neck lift and upper and lower eyelid lifts. With hospital and other fees, the total bill is about $4,000. This motive is also affected by what is covered by insurance or not (Landers, 2005; Moser, 2005). Since this customer motive will likely play a role for U.S. residents, it is discussed in a later section.

Timeliness Motive

A third motive is timeliness. The distinction between this motive and the availability motive is that the procedures are available in the home country, but there are long waiting lists. For example, MacIntosh (2004) mentions that across Canada, thousands are on waiting lists for surgeries that can last a year. This type of motive is likely to be found for residents of countries such as Canada, the United Kingdom, or Germany, but less likely to be found for U.S. residents since waiting lines are typically not that long in the United States (Hutchinson, 2005). The flow of these medical tourists is from countries with universal health care systems. The universal health care systems may cover much of the procedures but simultaneously create more demand for certain procedures or treatments that cause waiting lists. A country like the United States may become the destination for these tourists. In other instances, countries such as India may become the destination if the procedure can be provided there in a more timely manner (Gentleman, 2005).

Tourism Motive

The tourism motive combines any of the earlier motives with traveling to some exotic location for a vacation. For example, Rajagopal (2006) states that many of the tourists to Ayurveda in India combine their visit with a health care package. Most of these tourists end up in five-star spas and massage parlors, although there are also a few tourists who seek genuine treatment. Landers (2005) makes a similar observation when he states that for India, recovering patients often have time and money left for a visit to the Taj Mahal or another of India's sights. For this motive, there is also likely to be a move from a country such as the United States or Western European countries to other more exotic locations such as Costa Rica, India, or Thailand.

MEDICAL TOURISM AND THE UNITED STATES

Several motives for patients have been mentioned already. Why Americans may become medical tourists and the impact of this tourism on the U.S. health care system is reviewed.

Motives of Americans to Become a Medical Tourist

Americans may become medical tourists for different reasons. Availability of treatment, the cost of treatment, and timeliness of treatment are cited most often by patients (Hutchinson, 2005).

Availability. If treatment is not available in the United States, then alternative sites can be found easily with current availability of information such as the Internet. A variety of businesses in the travel and tourism industry have made it easy for a U.S. patient to shop and compare locations and prices.

Cost. For treatments that are available in the United States, some may be covered by insurance and some may not. With estimates of 43 million uninsured in the

Table 1. International cost comparisons in U.S.$. Based on (a) Nundy (2005), (b) Balaba (2005), (c) Firefly Medical Corporation, (d) Hutchinson (2005), (e) Reader's Digest (2006)

Procedure	India	Singapore	Thailand	United States
Coronary Angiogram	N.A.	2450 (e)	2886 (b)	25,6000 (e)
Coronary Bypass	10,000 (a)	N.A.	N.A.	86,000 (e)
Knee Replacement	16500 (c)	6207 (b)	7500 (b)	36300 (e)
Hip Replacement	16500 (c)	7200 (e)	N.A.	40800 (e)
LASIK Eye Surgery	730 (d)	730 (d)	750 (e)	4000 (e)

United States, coverage and insurance have no meaning for this group (Hutchinson, 2005). This means that cost becomes a major concern. Even for those who have insurance, the burden of coverage is shifting from the employer to the employee with higher deductibles and co-payments. As more plans become selective in what will be covered to keep premium prices down, even having insurance may cause a patient to look outside the United States because of cost.

In an interview with MedRetreat, currently the only U.S.-based business medical tourism company, MedRetreat indicated that the highest demand area for U.S. medical tourists is cosmetic/plastic surgery. Next is the demand for dental surgery, followed by orthopedic procedures. Cosmetic and dental procedures may not typically be covered by health insurance. MedRetreat has been in business for three years and has hospital partners in Argentina, Brazil, Costa Rica, India, Malaysia, South Africa, Thailand, and Turkey. All hospital partners are accredited through Joint Commission International or are in the process of accreditation. Hospitals are state of the art, and some are test sites for the newest technology being developed by companies such as General Electric, Philips, and Siemens.

MedRetreat stated that generally for most procedures, India will be the least expensive, and that in general the surgery procedures are 50% to 80% less expensive in their partner hospitals than in the United States.

Cost comparisons for types of medical procedures are provided in Table 1.

Timeliness. For many Americans, living with pain related to ambulation and mobility drives them to look for alternatives to free them from pain. Orthopedic surgeries most often are covered by insurance, but the criteria one has to meet before authorization is provided for this surgery can severely limit someone's life style. Also, medications used to alleviate pain can have long-term undesirable side-effects.

Tourism. Add to any of the aforementioned motives the opportunity for a luxury vacation paradise for recuperation with personal attendant care and the motivation to become a medical tourist increases. For example, a knee or hip replacement with a vacation thrown in has appeal because of the location and personal care provided during the time in the country. U.S. consumers are concerned about the typically short hospital stays that follow these surgeries and the type and place of follow-up care. A luxury setting (sometimes five-star facilities) with personal care, and a longer in-hospital recovery period at a reasonable price is much more appealing. Furthermore, even though the patient receiving the treatment may not feel up to a sightseeing vacation, those that are accompanying the patient may take advantage of it. For those patients who are ambulatory, a new country and location are incentives to exercise and see the sights, possibly speeding up the recovery process.

Destinations

The leaders in the medical tourism industry are Thailand, Singapore, India, and Costa Rica. Argentina, Brazil, Cuba, Jamaica, South Africa, Turkey, Jordan, Malaysia, Israel, Hungary, Latvia, and Estonia are all trying to break into the market (Hutchinson, 2005). The four leaders—Thailand, Singapore, India, and Costa Rica—are discussed.

Thailand. In a year-end briefing on this coming year's bright economic spots, Oliver (2005) states that 16 hospitals in Thailand are recommended for international patients. Bumrungrad Hospital in Bangkok stands out with accreditation by JCI. In 2004, more than 1 million medical tourists with 80% using private hospitals received medical care ranging from dental work to heart bypass surgery (Oliver, 2005). Oliver (2005) states, "At Bangkok's Bumrungrad Hospital, a standard cardiac surgical procedure costs on average US $7500, but it would cost $60,000 in a US hospital. Another benefit is that waiting lists are non-existent and patients enjoy luxury similar to a five-star hotel and personal care that would easily add thousands in a US situation. At US$ 110 million, the Bumrungrad ranks as one of the most technically advanced private hospitals in Asia. It is part of the Global Care Solutions network, a Luxembourg- registered company" (p.2). The hospital reported a 24% increase in patient volume in 2004 and is expecting a 15% increase in 2005 (Gentleman, 2005). Laohachewin (2005), the executive director of the Thai Trade Centre, states that the number of international patients has grown 200% in the last few years. A National Public Radio interview (Block, 2005) describes the hospital as a 17-story $100 million facility with a Starbuck's, McDonald's, and a sushi restaurant with 350,000 non-Thai patients. The hospital boasts of having more than 200 U.S. board-certified physicians. In 2005, Bumrungrad treated nearly 58,000 U.S. patients (Doheny, 2006).

Singapore. Singapore entered the medical tourism industry about a decade ago with the government banking on its clinical image as a key selling point in attracting 1 million patients by 2012 according to Oliver (2005). Singapore has a Johns Hopkins branch at the Singapore International Medical Centre. Singapore has nine Joint Commission International accredited hospitals. The Singapore Tourism Board has identified medical tourism as a high priority. Competition is strong for Singapore with Thailand and Malaysia. Singapore has generally higher costs than neighboring countries (Yong, 2004). In order to be competitive, Singapore has slashed fees by 20% to 30% in some areas (Brennan & Benzie, 2004). Loo Choon Yong, M.D., executive chairman of the Raffles Medical Group, states that Raffles Hospital, one of Singapore's top health care providers, was offering "prices that are competitive and comparable to the Thai hospitals" (Yong, 2004, p.889). To be competitive, the hospital was discounting 10 standard operations. According to Hutchinson (2005), more than 250,000 patients per year visit Singapore.

India. India is viewed as a relative newcomer in the industry (Hutchinson, 2005). With strong government backing, India hopes to be the destination point for medical tourism. The Ministry of Tourism is working on a campaign to be launched oversees (Chowdhury, 2005). Additionally, the government has set up several expert committees to deal with medical insurance and human resource development. Currently, India has two hospitals that are Joint Commission International accredited. Medical tourism is "forecast to become a $2.3 billion business for India by 2012, and some analysts predict that it will be the next major driver of the Indian economy after information technology" (Gentleman, 2005, p. 1). Already, about 150,000 visitors seek treatment every year, and the figure is rising at 15% annually (Gentleman, 2005). By marketing India as a global medical tourism destination, the Indian government hopes to capitalize on the low-cost, high-quality medical care available in the country (Businessline, 2005).

India has some clear advantages over other countries. Their educational system is producing many medical professionals, and some hospitals already have international reputations. It has kept up with technological advances, and price is a major selling point. Some of their hospitals use telemedicine to communicate with locations internationally, and teleconferencing enables them to follow up with their patients around the world. "Unlike many of its competitors, India also has the technological sophistication and infrastructure to maintain its market niche. Also, Indian pharmaceuticals meet the stringent requirements of the U.S. Food and Drug administration. Additionally, India's quality of care is up to American standards, and some Indian medical centers even provide services that are uncommon elsewhere" (Hutchinson, 2005, p. 4).

Costa Rica. For the North American patient, the Asian destinations are considered very far away because of the flight distance. Costa Rica provides a South American location that is known for its inexpensive, high-quality medical care. This country is noted for its commitment to sanitation and high-quality medical care for its citizens, which makes it a great location for Westerners. Costa Rica has no JCI-accredited hospitals. A 1991 survey conducted by the University of Costa Rica found that 14.25% of all visitors to Costa Rica come to receive some sort of medical care (Underwood, 1996). Costa Rica is renowned for its plastic surgery services. While this country has a reputation for plastic surgery, it should be noted that a full range of medical procedures is available, including transplants. Physicians are mostly educated in Costa Rica and then branch out around the globe to enhance their education, with many speaking multiple languages (Underwood, 1996).

Two well-known private hospitals, Clinica Biblica and Clinica Catolica, offer first-class, ultra-modern services, lower than U.S. costs, and affiliations with U.S. hospitals (Underwood, 1996).

Selection of a Destination

In the previous sections, the emphasis has been on the availability and/or costs. When medical tourists choose an international location for their medial treatment, they should also be concerned with the quality of the treatment. Aside from personal care and a luxurious room, how do these foreign locations rate? Essentially, patient safety is the number-one concern. Factors that are related to patient safety are the education and credentials of the staff, starting with the physician. Another factor is accreditation. If a hospital has the hospital JCI accreditation, then it assures that the hospital meets the standards that most U.S. hospitals meet.

Another important issue relevant for the U.S. health care system is the amount of time for recovery for the procedure. What is, for example, the follow-up and by whom? This includes the type of professional, but also whether, for example, the follow-up has to be conducted in the United States since this may place an unfair burden on the U.S. hospital system.

Yet another issue is whether the overseas hospital has electronic copies of the patient files. From a patient safety and continuity of care perspective, electronic medical records provide a means of communication from other countries to the United States. Medications, follow up care, and communication from the primary physician to the specialist in another country can occur.

FUTURE TRENDS: IMPACT ON THE U.S. HEALTH CARE INDUSTRY

Loss of income. The international lure for less expensive procedures is driving cash customers for health care out of the United States for both uncovered as well as (partly) covered services. For example, the Bumrungrad Hospital states that it served 58,000 Americans in Thailand in 2005 (Doheny, 2006). Consequently, the income generated from these U.S. patients is lost for the U.S. hospitals and the U.S. economy. The cumulative amount of money spent by U.S. medical tourists elsewhere may be a significant loss for the U.S. economy. Currently, most of the U.S. medical tourists pay for the medical treatments themselves, because it is either not covered by their insurance or only partly covered, which still makes it cheaper to go overseas. One can only speculate what would happen if U.S. health insurance companies realize the cost savings they can achieve by sending insured patients to cheaper overseas hospitals.

Cost to U.S. hospitals. Why are the costs in U.S. hospitals so high? As a parallel, in manufacturing situations, low labor-cost countries are also often thought of as attractive, but although the labor cost may be low (in cost per hour), the total cost of manufacturing goods in other countries may turn out to be much higher than in the United States (Markides & Berg, 1988). This is the case because the productivity of labor is often lower, and more importantly, labor is usually only a small percentage of the total cost (on average 15%). Other costs involved include costs related to quality, materials, machinery, and so forth, which are often higher in low labor-cost countries. For hospitals, this is not the case (i.e., labor costs are the major costs for a hospital (Finkler, 2005). Therefore, the low labor-cost countries have a major advantage that is not likely to disappear in the near future.

U.S. hospitals are at best in a fragile state (American Hospital Association, 2005). Continued vacancies for nurses, pharmacists, and imaging and lab technicians are affecting patient access to services. Work force shortages are projected to get worse, not better. Rising costs of pharmaceuticals and other medical supplies put a further strain on hospital budgets. Payment shortfalls and the growing number of uninsured are pressuring the current hospital systems. The costs of the U.S. health care system are high. The trend of U.S. medical tourists may make it worse. If fewer patients are served, then the fixed costs of U.S. hospitals have to be spread among fewer patients, which increases the cost for those patients. This in turn makes medical tourism become even more attractive. Such a cycle could spiral out of control with the result that only emergency treatments or procedures, or treatments that have to take place locally will remain in the United States.

Competition. U.S. hospitals may now face international competition. This is essentially an alien concept for hospitals that traditionally have a geographically determined patient base. Now this situation is changing, which impacts the income generated in U.S. hospitals. To make matters worse, U.S. hospitals may not only be competing with international health care systems. Many of the nations' governments mentioned earlier have targeted medical tourists explicitly. For example, a company like MedRetreat is clear about pointing out that over the last 10 years, global advances have been made in countries that desire to build their economy. Health care has provided a means for these countries to look at attracting the relatively wealthy U.S. patient. The ease of access to information on the Internet has helped to inform the U.S. consumer. Responses for organizations like MedRetreat are immediate and very customer friendly. Aside from this government support, these countries also are explicitly targeting tourists. Their tourist offices are involved in attracting medical patients. This is a completely new way of competing. More research needs to be completed regarding U.S. insurance and self-insured businesses' view of medical tourism.

CONCLUSION

This chapter has explored the issue of medical tourism from a U.S. perspective. The motives for patients to become medical tourists and the potential destinations for these patients have been described. The consequences for the U.S. health care system have been examined. At the moment, it seems that the U.S. health care system is extremely uncompetitive and unprepared to deal with this new international challenge.

REFERENCES

American Hospital Association. (2005). *Taking the pulse: The state of America's hospitals*. Retrieved March 9, 2006, from http://www.aha.org

Balaba, R. (2005). RP needs to play catch-up in medical tourism game. *Business World*. Manila. Retrieved February 25, 2006, from ProQuest database.

Block, M. (2005). All things considered: Profile & increasing popularity of medical tourism. *National Public Radio*. Retrieved February 25, 2006, from ProQuest database.

Bradley, R., & Kim, E. (1994). Loosening the reins: Autonomy boosts Cuban medical industry. *Harvard International Review, 16*(4), 66–70.

Brennan, R., & Benzie, R. (2004). Countries offer quick getaway from waiting lists: Canadians must pay high prices, but medical tourism taking off. Orthopedic clinics advertise fast service, attractive surroundings. [ONT Edition]. *Toronto Star*. Retrieved February 25, 2006, from ProQuest database.

Businessline, Chennai. (2005). Government plans to promote medical tourism from Oct. text. Retrieved February 25, 2006, from ProQuest database.

CBC News Online. (2004). Medical tourism: Need surgery, will travel, part 1. Retrieved March 12, 2006, from http://www.cbc.ca/news/background/healthcare/medicaltourism.html

Chowdhury, R. (2005). Medical tourism to attract 1 m visitors by 2010. Businessline, Chennai. Retrieved February 25, 2006, from ProQuest database.

Doheny, K. (2006). Healthy traveler; a little sightseeing, a little face-lift; medical tourism takes wing as Americans seek out vacation-surgery getaways. *Los Angeles Times*. Retrieved February 25, 2006, from ProQuest database.

English, V., Mussell, R., Sheather, J., & Sommerville, A. (2005). More on medical tourism. *Journal of Medical Ethics*, *31*(12), 743.

Finkler, S. (2005). Cost containment. In A.R. Kovner, & J.R. Knickman (Eds.), *Jonas & Kovner's* h*ealthcare delivery in the United States*. New York: Springer.

Firefly Medical Corporation. Retrieved April 18, 2006, from http://www.firefly.com

Gentleman, A. (2005). Lines drawn in India over medical tourism. As business booms, some doctors see domestic patients getting shortchanged. *International Herald Tribune.* Retrieved February 25, 2006, from ProQuest database.

Hutchinson, B. (2005). Medical tourism growing worldwide. *University of Delaware U Daily*. Joint Commission International, JCI Accredited Organizations. Retrieved April 13, 2006, from http://jointcommissioninternational.com.

Landers, J. (2005). Americans visit India for high-quality "medical tourism." *Knight Ridder Tribune Business News*. Retrieved February 25, 2006, from ProQuest database.

Laohachewin, S. (2005. 200 percent increase in Thailand's medical tourism attributed to Middle East visitors. *Middle East Company News*, Dubai. Retrieved February 25, 2006, from ProQuest database.

MacIntosh, C. (2004). Medical tourism: Need surgery, will travel, part 2. CBC News Online. Retrieved February 14, 2006, from http://www.cbc.ca/news/background/healthcare/medicaltourism.html

Markides, C.C., & Berg, N. (1988). Manufacturing offshore is bad business: To stay competitive, stay home. *Harvard Business Review, 10*, 113–120.

MedRetreat. Retrieved February 25, 2006, from http://www.medretreat.com.

Moser, L. (2005). The medical tourist: How my shoulder sent me to China. Retrieved April 5, 2006, from http://www.slate.comid/2131609

Nundy, S. (2005). Lines drawn in India over medical tourism. As business booms, some doctors see domestic patients getting shortchanged [IR Edition]. *International Herald Tribune*. Retrieved February 25, 2006, from ProQuest database.

Oliver, C. (2005). Going under the knife never looked better. Medical tourism is booming, but managers are caught out. *South China Morning Post*, Hong Kong. Retrieved February 25, 2006, from ProQuest database.

Rajagopal, S. (2006). Promoting medical tourism. *The Hindu*. Retrieved February 25, 2006, from ProQuest database.

Reader's Digest Special Report. (2006). I can't afford to get sick.

Underwood, E. (1996). *To your health in Costa Rica*. Retrieved April 5, 2006, from http://222.cocori.com/library/life/med1.htm

Xinhua News Agency. (2005). Cyprus plans to tap medical tourism. *Xinhua News Agency–CEIS*. Retrieved February 25, 2006, from ProQuest database.

Yong, L. (2004). Medical tourism: Singapore seeks to challenge Thailand as hotspot for medical tourism. *Life Science Weekly*. Atlanta. Retrieved February 25, 2006, from ProQuest database.

KEY TERMS

Asian Health Care: Those health care organizations and systems that provide health services in Asia and the Pacific Rim.

Health Care Competition: Rivalry between health care organizations and health care systems for patients.

Health Care Insurance: The contractual relationship that exists to provide health care services between the insurer and the insured.

Hospitals: Institutions providing medical or surgical care and treatment for the sick and injured.

Medical Tourism: A need for a medical procedure or treatment in combination with a motive of availability, timeliness, or cost with the desire to travel to another country where the individual chooses another country for care.

Medical Tourism Companies: Companies assisting patients to have a medical procedure or treatment in another country.

U.S. Health Care Industry: The U.S. health care system is comprised of for profit, not for profit, and governmental health care organizations and systems.

A Model to Evaluate the Flow of an Emergency Department

M.C.J. Schnellen
University of Twente, The Netherlands

INTRODUCTION

Little is published about generic models that describe the physic and information process of an emergency department. This chapter expounds the development of a general model that describes the current physic and information processes of an emergency department; it is based on observations of emergency departments in The Netherlands and the United States. This model can be of great value for the management of an emergency department, since it is useful to first analyze a current situation before improving organizational processes (Daft, 2000).

The main purpose of this chapter is formulated as "the creation of a model that can evaluate the physic and information flow of an emergency department by using literature and emergency department observations." This research investigates if it is possible to develop this kind of model, and if this is possible, then how this model can be used to compare the physic and information flows for different emergency departments.

This research is defined inside one of the projects of the department Organizations, Operations and Human resources of the University of Twente. This is a four-year project involving the emergency department of a large hospital in The Netherlands, led by ir. R. J. Rosmulder and guided by prof. dr. ir. J.J. Krabbendam. The project includes the improvement of organizational processes at this emergency department by using business theories in order to develop new ways of controlling the organizational processes of the emergency department.

BACKGROUND

In order to develop the model, observations were made in the emergency departments of larger hospitals in The Netherlands and the United States. The United States was chosen because many differences were expected between the Dutch and American emergency departments due to the more denationalized health care, the managed care, and the differences in medical education in the United States.

The foundation of the developed model is based on Miller and Rice (1967) who developed a theory of one generic framework where organization, human activities, and tasks are conciliated. Based on the findings of Miller and Rice (1967), the developed model encloses tasks, human activities, and organizational aspects. In addition, the operations model of Slack, et al. (1998) is used. Here, Slack, et al. (1998) describe the division in design of the process, planning, and control of the process and improvement of the process. Since only the current process of an emergency department is involved in the model, it demarcated the model boundaries. In addition, to gain knowledge about the input, throughput, and output of the process, and especially about the input and output of patients, the model of Asplin (2003) is used. To use a framework in this research, the theory mentioned in Processen in Organisaties is used, written by dr. ir. H. Boer and prof. dr. ir. J.J. Krabbendam (1996). They expose a step-by-step plan to analyze the different characteristics of a process inside an organization. To analyze the two observed emergency departments, their division is used.

DEVELOPING THE MODEL

Simplification of the Model

Similar to manufacturing organizations as referred to in Slack (1998), in an emergency department there is the input, throughput, and output. The input contents patients who enter the emergency department. These patients can be seriously ill or just need some care on that specific moment. Examples of different inputs are given by Asplin (2003).

The throughput is divided into triage, making a diagnosis, treatment inside the emergency department, and a referral to a place outside the emergency department. As heard from authorities and seen in hospitals and emergency department environments, this is an accepted approach of medical personnel to diagnose and treat patients. In this figure, it is assumed that every patient that arrives at the emergency department will successively pass through all these steps except triage. As mentioned in Ceglowski, Chirilov, and Wassertheil (2005), the emergency process is hard to describe, for example, because one nurse serves many patients. In this model, the patient is the central part; it is not mentioned what percentage of the nurse's time is spent on that particular patient. The patient flows through the process.

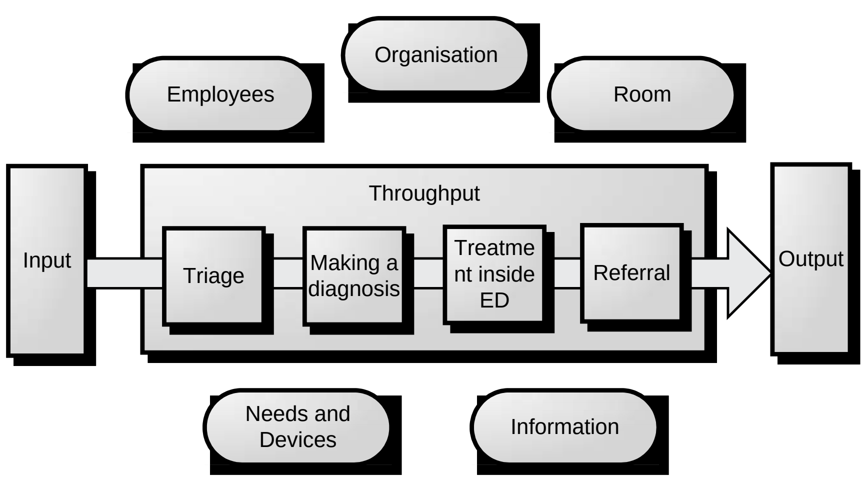
During the throughput process, the organizational arrangements like structure, culture, and procedures are of great influence (dr. ir. H. Boer and prof. dr. ir. J.J. Krabbendam, 1996). In addition, the amount and kind of employees, rooms, needs, devices, and information change. In the triage step of the throughput, for example, there is most of the time one nurse needed, a triage room or waiting area, some simple medical instruments, and information like the patient's medical history. For every step (triage, making a diagnosis, treatment inside the emergency department, and the referral) a combination of employees, rooms, needs, devices, and information is used during that part of the process. By describing every step, at the end the patient flow, its needs are described. In Figure 1, the simplification of the model is presented.

In order to present a simplified model, very specific flows could not be shown here. The output of this process could be cured patients, but far more is possible. Patients can be sent to their homes, admitted to the hospital for either surgery or nurse ward, or referred to another care institute. Asplin (2003) mentioned some in his model, and after experiences in emergency departments, there are some more mentioned in the next model.

The Extended Model

This model has the same basis as the simplification presented previously, but in a more extensive framework and, therefore, more complete. The

Figure 1. The simplification of the model



model is, as mentioned, a framework for the primary physic and information flow process. According to de Vries, Bertrand, and Vissers (1999), to analyze this process, the input must be divided into groups. The division for the input of patients is made by fast-track patients, urgent care patients, and trauma patients. Fast-track patients are patients who don't need urgent care and don't need devices like monitors in their treatment room. Examples of these patients are patients with a simple broken wrist or who need to be sutured for a small wound. Urgent patients in this rapport are defined by patients who need monitoring and specialized care, which will be determined by triage. Trauma patients in this rapport are patients who need immediate care, monitoring, and multiple specialists. This group doesn't see the triage at all, but goes directly to the step where the specialist begins making a diagnosis. The arrows of the first two inputs are already pointing at the second step of the primary process. Take, for example, the step Treatment inside the emergency department. A patient needs an emergency department doctor, a specialist, or a nurse, depending on the group into which the patient is divided. In addition, for this patient, an emergency department treatment room or sometimes the special plaster room can be used. Needs and devices in this step can be medicines or plaster casts or suturing needs. Information needed in this step is, for example, the diagnosis information from the previous step throughput, which is marked in pink. The organization influences the process of the patient through his or her structure and culture. The visualization of this model is displayed in Appendix 1.

FUTURE TRENDS

This model is an example of using management techniques in health care. Management techniques are already applied in health care, but many more applications are possible. The implementation of technologies widely used in other businesses, like operation models or product tracking systems, can be very useful in healthcare as well. Research in healthcare management will grow and become more important. Like other industries, in health care there is a large amount of money presented. Consequently, cost savings due to implementation of management techniques will become more attractive.

An important note is the need to change these business models to applicable models for healthcare, since health care is not comparable with other industries. Management techniques designed for business industries should be adjusted to the characteristics and needs of healthcare; otherwise, there will be a large risk of failure with implementation of these techniques.

CONCLUSION

The model is applied in two emergency departments and proved to be successful. It can be used as guidance for evaluating the physic and information flow in a structural way. The model turned out to provide the user with a lot of useful information about the organization of an emergency department. Using this model for a comparison between emergency departments is more difficult, since the patients flow in detail differ too much, and consequently, different data are gained for every emergency department. In addition, a comprehensive model is complete enough to use for this purpose. But unfortunately, the author did not succeed in developing a proper and complete simplification of this model. Although it would be easy to use a simplification of the model, the level of this simplification is too high to use, since some important steps cannot be visualized.

Hence, this model is a proper tool to evaluate the current physic and information flow of one emergency department, but it did not prove itself as a good way to compare emergency departments, especially due to the totally different ways of working, different inputs of patients, and different environments.

REFERENCES

Academisch Medisch Cetrum Amsterdam. (2006). Retrieved June 22, 2006 through September 14, 2006, from http://www.amc.nl

Asplin, B.R., et al. (2003). A conceptual model of emergency department crowding. *Ann Emerg. Med, 42*, 173–180.

Boer, dr. ir. H. Krabbendam, prof. dr. ir. J.J. (1996). Processen in organisaties. *Reader of mens, technologie en organisatie, I*, 17–48.

Ceglowski, R., Churilov, L., & Wassertheil, J. (2005). Don't panic—Prepare: Towards crisis-aware models of emergency department operations. *Clin Invest Med, 28*(6), 320–322.

Daft, R.L. (2000). Management. Fort Worth, TX: The Dryden Press.

de Vries, G., Bertrand, J.W.M., & Vissers, J.M.H. (1999). Design requirements for health care production control systems. *Production Planning & Control, 10*(6), 559–569.

Elsevier. (2006). Beknopte gids voor het nieuwe zorgstelsel. Retrieved July 12, 2006, from http://www.elsevier.nl/nieuws/nederland/nieuwsbericht/asp/artnr/76852/index.html

Miller, E.J., & Rice, A.K. (1967). Systems of organization: The control of task and sentient boundaries. London: Tavistock Publications.

Mintzberg, H. (1997). Toward healthier hospitals. *Health Care Management Review, 22*, 9–18.

Mount Sinai Hospital. (2002). http://www.sinai.org/

Peppard, J., & Rowland, P. (1995). The essence of business process re-engineering. *Reader of Productiemanagement*, 2, 45–46.

Rosmulder, R.J. (2005)

Schnellen, M.C.J. (2006). Developing a model to compare two emergency departments in The Netherlands and The United States.

Slack, N., et al. (1998). *Operations management*. London, Financial Times.

Steehouder, M., et al. (1999). Leren communiceren. Groningen: Wolters Nordhoff.

Storm-Versloot, M. (2005).

van der Bij, J.D., et al. (1999). Improvement and renewal of health care processes: Results of an empirical research project. Health Policy, 48, 135–152.

van Katwijk, J.J. (2001). An exploration of future hospital care in 10 countries. Reader international health strategy and hospital of the future. Amsterdam: Public Space.

Verschuren, P., & Doorewaard, H. (2005). *Het ontwerpen van een onderzoek*. Meppel: Boom/Lemma.

Wikipedia. (2006). Eerstelijns gezondheidszorg. Retrieved July 13, 2006, from http://nl.wikipedia.org/wiki/Eerstelijns_gezondheidszorg

World Health Organization (2006). Country information. Retrieved July 13, 2006, from http://www.who.int/countries/nld/en/ and http://www.who.int/countries/usa/en/

Zorgverzekeraars Nederland. Retrieved August 3,, 2006, from http://www.zn.nl

Zorgverzekeraars.nl. (2006). Het nieuwe zorgstelsel. Retrieved July 12, 2006, from http://www.zorgverzekeraars.nl/feit_nieuwstelsel.html

KEY TERMS

Emergency Department: A department within a hospital where immediate urgent care is delivered to patients needing noncritical, urgent, and trauma care.

Emergency Room: The space within an emergency room where immediate urgent care is delivered to patients needing urgent and trauma care.

Fast-Track Patients: Are patients who don't need urgent care and don't need devices like monitors in their treatment rooms.

Healthcare: The service that treats and promotes the well-being of a human being. It contains, but is not limited to, prevention, diagnostics, treatment, and therapeutic and palliative care.

Information Flow: The flow of the information needed during the input-throughput-output process.

Physic Flow: The flow of the patient and the employees, devices, and rooms needed during the input-throughput-output process.

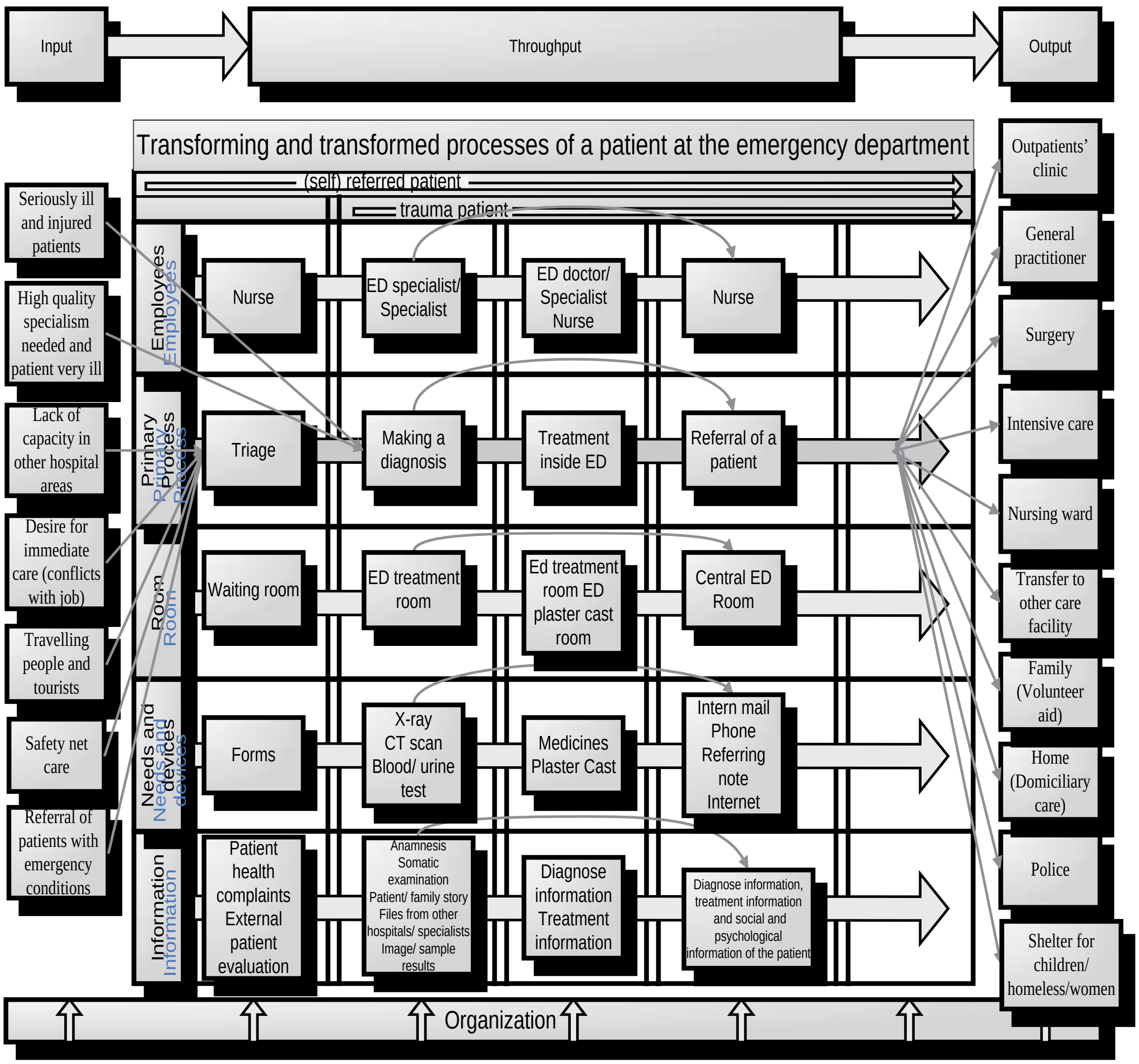
Process Model: A well-known term in process engineering. It is a model that shows the process that can help, but is not limited to evaluating or (re-)designing a process.

Trauma Patients: Patients who need immediate care, monitoring, and multiple specialists. This group doesn't see the triage at all.

Urgent Patients: Patients who need monitoring and specialized care, which will be determined by triage.

APPENDIX 1: THE EXTENSIVE MODEL

Figure A1. The extensive model



Modification of Arruda's Accessory Pathway Localization Method to Improve the Performance of WPW Syndrome Interventions

Sándor Miklós Szilágyi
Hungarian Science University of Transylvania, Romania

László Szilágyi
Hungarian Science University of Transylvania, Romania

Constantin T. Luca
University of Medicine and Pharmacy, Timişoara, Romania

Dragos Cozma
University of Medicine and Pharmacy, Timişoara, Romania

Gabriel Ivanica
University of Medicine and Pharmacy, Timişoara, Romania

Zoltán Benyó
Budapest University of Technology and Economics, Hungary

INTRODUCTION

Description of the Wolff-Parkinson-White Syndrome

The Wolff-Parkinson-White (WPW) syndrome is characterized by an accessory pathway (by-pass tract) between the atria and ventricles that conducts parallel with the atrioventricular (AV) node-His bundle, but faster (Wolff, Parkinson, & White, 1930; Yee, Klein, & Guiraudon, 1995). An accessory AV connection can conduct in both directions. The presence of these by-pass tracts may predispose to atria-ventricular reentrant tachycardia. Moreover, in the setting of atrial fibrillation, the WPW syndrome may cause a catastrophically rapid ventricular response with degeneration to ventricular fibrillation (VF).

Electrocardiographically, the WPW syndrome can be characterized by a specific sinus rhythm pattern. Its other specific features are paroxysms of re-entry tachycardia (the incidence in the young adult population is about 10% and growing up with age to 30%), more rarely paroxysm of atrial fibrillation (20-30% of patients with the syndrome), or atrial flutter (Guize, Soria, Chaouat et al., 1985; Wellens, Fare, & Bar, 1987).

In the case of WPW syndrome, the electrocardiogram (ECG) tracing is a mixture of the electrical activities caused by the accessory AV connection and normal AV conduction system. The fast impulse conduction produces an initial deflection in the QRS complex (delta wave).

The length of this delta wave is determined by the difference between the accessory AV connection and normal AV conduction times. The modified ventricular activation may cause secondary abnormalities in the ventricular repolarization such as: ST segment displacement (elevation or depression), T wave shape distortion, and abnormal U wave appearance. The conduction capacity variances (changes may occur hour by hour or day by day) of the accessory AV connection may result in alternating WPW pattern (complete, partial, or missing pre-excitation, concertina effect).

An adequate analysis of this phenomenon is important, since 1-2‰ of the population suffer from WPW syndrome (Wellens, Brugada, Penn et al., 1990). When the refractory period of the accessory connection becomes too short, the patient's life is endangered by a possible VF. Unfortunately, the exact risk for developing VF during high ventricular rates is unknown.

WPW Syndrome Analysis

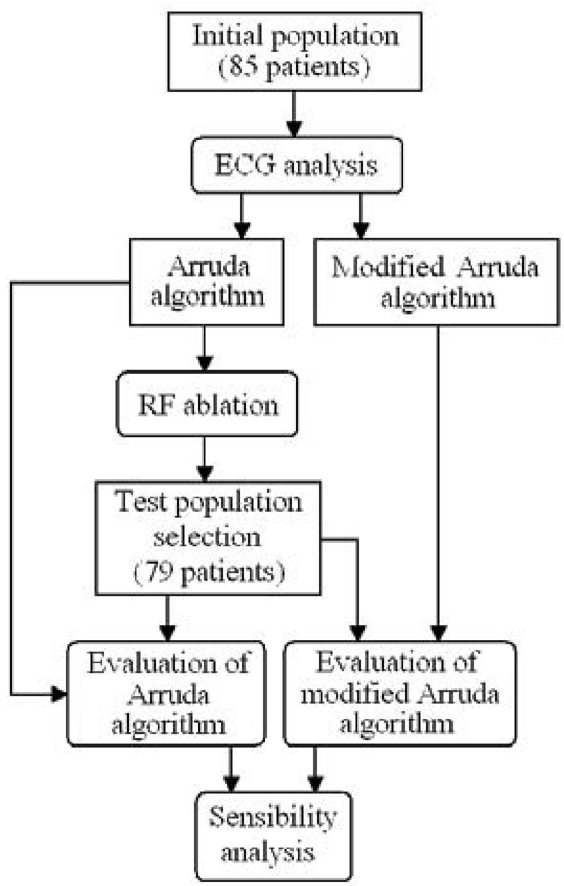
Usually the WPW analysis is focused to develop and validate an accessory pathway (AcP) localization method (Rosenbaum, Hecht, Wilson, & Johnston, 1945). A number of investigations have correlated ECG patterns and processing algorithms for detecting the place of the AcP (Arruda, McClelland, Wang, Beckman et al., 1998; Fitzpatrick, Gonzales, Lesh et al., 1994; Reddy & Schamroth, 1987), while other studies have been focused on the localization, realized through three-dimensional (3D)-heart reconstruction by the inverse solution of the ECG (Huiskamp & Greensite, 1997).

Unfortunately, the inverse problem, in contrast to the forward approximation methods, does not possess a mathematically unique solution. Another not easily by-passable difficulty is rooted in its ill-posed nature, whereby the obtained inverse solution could be unstable and may oscillate widely with the slightest perturbation.

Several approaches have been explored to handle the problem of multiple solutions by using equivalent cardiac generators, such as equivalent dipole (De Guise et al., 1985) and multipole, heart surface isochrones (Huiskamp & Greensite, 1997), or epicardial potential (Guanglin & Bin, 2001; Shahidi, Savard, & Nadeau, 1994). The high sensitivity of solutions to the different disturbances forced the investigators to explore various regularization techniques (Shahidi et al., 1994).

These methods allow a significant progress, but the diverse uncertainty elements of the processing limit

Figure 1. Flow sheet of the research



the potentially beneficial ECG inverse solutions from becoming a routine clinical tool at present.

The Arruda method differs from prior algorithms in its combined use of the resting ECG, utilization of only five ECG leads (I, II, III, V1, aVF), and by prospective validation of the algorithm.

In this article, we present a sensibility analysis of the Arruda's stepwise method (Arruda et al., 1998), and a decomposition algorithm to improve the performance of AcP localization. Our main purpose is to detect more accurately the location of the ventricular insertion.

MATERIALS AND METHODS

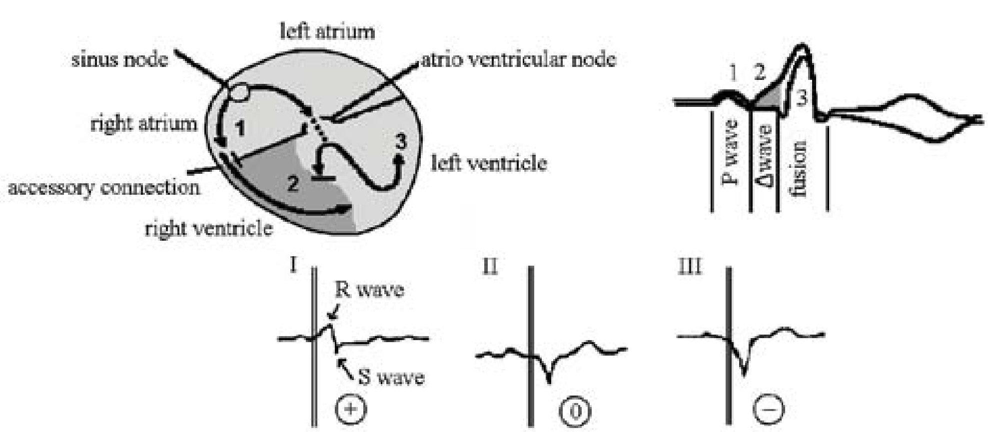
Base Studies

This research starts from the results and considerations published in S. M. Szilágyi's "Sensibility Analysis of the Arruda Localization Method" article (Szilágyi, Benyó, & Frigy, 2004). It has been concluded that the efficiency of the localization method—described in Arruda's "Development and Validation of an ECG Algorithm for Identifying Accessory Pathway Ablation Site in Wolff-Parkinson-White Syndrome" work (Arruda et al., 1998)—depends significantly on the location of the AcP.

According to the above-mentioned studies, we developed a modified Arruda localization method, which was tested and evaluated through our measurements. The schematic representation of the research process can be seen on Figure 1.

The *initial population* (1) of the research consisted of *85 patients* having at least one AcP (delta wave was detected in the preliminary phase of the study). For each patient a 5-10 minutes long 12-lead *ECG* recording was registered. The so obtained data were then preprocessed and *analyzed* (2) in order to provide input for stages 3-4. Thereafter, we separately predicted the AcP locations by means of the *Arruda* (3) and *modified Arruda* (4) *algorithms*. After *radio-frequency (RF) catheter ablation* (5), 6 patients were excluded,due to multiple AcP's, while the remaining *79* were selected as *test population* (6) for further study phases. In the course of *evaluation of Arruda* (7) *and modified Arruda* (8) *algorithms*, the estimated positions of both localization methods (3-4) were evaluated one by one and validated with the outcome of *RF ablation* (5). Finally, within the frame of the *sensibility analysis* (9), the results of stages 7-8 were statistically compared, followed by a vector-space dissection with regard to the spatial location of the ECG leads used by each algorithm.

Figure 2. The emergence of WPW syndrome; the delta wave polarity



Initial Arrangements

In the preliminary phase of the study, we performed a 12-lead ECG measurement for patients manifesting cardiac symptoms that served as a basis to select the subjects suffering from WPW syndrome (initial population—85 persons). The ECG recordings were sampled at 500 Hz frequency with 12-bit resolution.

Our ECG signal processing concerned only accurate detection of the QRS complexes and delta (Δ) waves presented in Figure 2, constituting input parameters for the Arruda and modified Arruda algorithms.

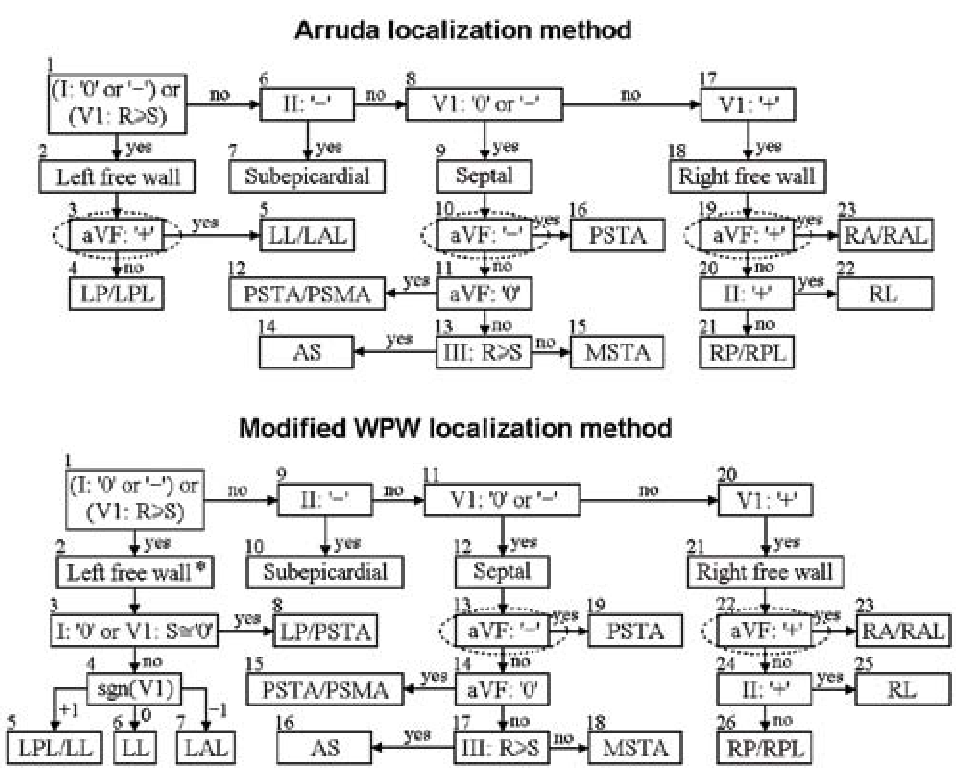
As Figure 2 presents, one cause for a wide QRS-complex that exceeds 0.12 seconds may be the Wolff-Parkinson-White syndrome (WPW syndrome). In this case the QRS-complex initially exhibits an early upstroke called *delta wave*. The two presented ECG signals differ significantly in shape. In both cases the interval from the P-wave to the R spike is normal, but the early ventricular excitation forming the delta wave shortens the PQ-time.

The cause of the WPW syndrome is the transfer of activation from the atrium (region 1 in Figure 2) directly to the ventricular muscle via an abnormal route, called AcP, which bypasses the AV junctions. This partially activates the ventricular muscle (region 2 in Figure 2) before normal activation (region 3 in Figure 2) reaches it via the normal conduction system (after a delay in the AV junction). This process is called pre-excitation, and the resulting ECG depends on the specific location of the accessory pathway (see right side of Figure 2). With the help of localization methods the place of AcP can be determined.

The Arruda Algorithm

The Arruda algorithm is a well known, clinically tested localization method, developed in order to identify the place of a singular accessory connection, reaching up to 90% recognition rate. It uses only five leads (I, II, III, aVF, V1) from the 12-lead ECG recordings.

Figure 3. Presentation of the Arruda and modified Arruda localization method; the left bottom side shows the modified left ventricular branch of the modified algorithm



The first step of the algorithm is to analyze the amplitude relations of the R, S, and delta (Δ) waves in order to determine the AcP location (see Figure 2). The onset of the delta wave in each lead is measured from the onset of the earliest delta wave in any of the 12 leads. After 20 ms the displacement of the delta wave in each lead is classified as positive (+), negative (–) or isoelectric (0) (see Figure 2).

The Arruda algorithm specifies a heart-region as the location of the AcP. According to Arruda, the possible AcP places have been divided into three major regions, which are further divided thereafter, as follows:

- Septal accessory pathways: anteroseptal tricuspid annulus and right anterior paraseptal (AS/RAPS), mid-septal tricuspid annulus (MSTA), posteroseptal tricuspid annulus (PSTA), posteroseptal mitral annulus (PSMA), subepicardial posteroseptal (SEC);
- Right free-wall accessory pathways: right anterior (RA), right anterolateral (RAL), right lateral (RL), right posterolateral (RPL), right posterior (RP);
- Left free-wall accessory pathways: left anterolateral (LAL), left lateral (LL), left posterolateral (LPL), left posterior (LP).

The proper Arruda algorithm might be interpreted as a decision tree consisting of simple two-way (YES/NO) ramifications, so called decision points (see Figure 3). It has five leads (I, II, III, aVF, V1), R and S waves, respectively the displacement of the delta wave in each lead (+, -, 0) as input, while the above mentioned heart-regions (locations) as output.

The Modified Arruda Algorithm

Although it has been statistically concluded that the performance of the Arruda algorithm is about 90%, some modifications in this localization method could be beneficial.

In Szilágyi et al. (2004), we pointed out that in most of the cases the estimation errors of the Arruda algorithm are correlated with aVF sign test. In the same place, we suggested modifying the lead aVF where it is possible.

The modified Arruda algorithm (see Figure 3) developed in order to identify the place of a singular accessory connection uses the same leads, same heart regions for localization of the AcP and same displacement of the delta wave in each lead (+, -, 0) as Arruda. The algorithm might be interpreted as a decision tree consisting of simple two-way (yes/no) ramifications with one exception at place 4 (see Figure 3). Although Szilágyi et al. (2004) revealed that the most sensible aVF sign tests at locations 3, 10, and 19, in Figure 3 we modified only the first one at left free wall (see Figure 3). The other two sensible places are marked as places 13 and 22 in Figure 3. The left free wall region is completely altered for better performance but all other tests are the same as in Arruda's work.

Radio-Frequency Catheter Ablation

RF catheter ablation of cardiac tissue is highly effective in the invasive treatment of WPW syndrome (Cao et al., 2000). During RF cardiac ablation, a catheter is introduced into a heart chamber via percutaneous peripheral venous or arterial conduits and placed in contact with the target ablation region (such as an AcP) at the endocardial surface. A high frequency current is applied between the catheter electrode and a dispersive electrode attached to the patient's skin. The myocardium is heated by Joule heat and heat conduction inside the myocardium. A temperature of 50 C° or higher causes irreversible loss of cellular excitability of tissue and forms a lesion.

In the course of RF ablation, a priori knowledge about the target ablation site(s) could be beneficial due to shorter intervention time and therefore lower risk. The Arruda (as well as the modified Arruda) algorithm predicts the location of a singular AcP, serving as a starting point to the surgeon. However, the exact number and place of the AcPs will be determined only during the surgical operation. Therefore, RF ablation plays a validation role for both localization methods. By means of RF ablation, we excluded from the research 6 patients with multiple AcPs, thus obtaining a test population consisting of 79 subjects.

Table 1. Detection accuracy of the accessory pathway with Arruda and modified Arruda localization method

Ablation Site	Number	RA/RAL	RL	RP/ RPL	AS/RASP	MSTA	PSTA	PSTA/PSMA	LP/LPL	LL/ LAL	SEC	Sensitivity (%)	Specificity (%)
RA/RAL	11	11										100.0	97.1
RL	8	1	7									87.5	100.0
RP/RPL	6	1		5								83.3	100.0
AS/RASP	2				2							100.0	100.0
MSTA	3					3						100.0	98.7
PSTA	14					1	12		1			85.7	100.0
PSTA/PSMA	1							1				100.0	98.7
LP/LPL	2							1	1			50.0	96.2
LL/LAL	22								2	20		90.9	100.0
SEC	10										10	100.0	100.0
All	79											91.1	99.0

Ablation Site	Number	RA/RAL	RL	RP/ RPL	AS/RASP	MSTA	PSTA	PSTA/PSMA	LPL/LL	LL	LAL	LP/PSTA	SEC	Sensitivity (%)	Specificity (%)
RA/RAL	11	11												100.0	97.1
RL	8	1	7											87.5	100.0
RP/RPL	6	1		5										83.3	100.0
AS/RASP	2				2									100.0	100.0
MSTA	3					3								100.0	98.7
PSTA	14					1	13							92.9	100.0
PSTA/PSMA	1							1						100.0	100.0
LPL/LL	1								1					100.0	98.7
LL	10								1	9				90.0	100.0
LAL	12										12			100.0	100.0
LP/PSTA	1											1		100.0	100.0
SEC	10												10	100.0	100.0
All	79													94.9	99.5

Table 2. Global accuracy of each decision point for Arruda and modified Arruda localization method

Place nr.	Total decisions	Failed decisions	Performance
1	79	1	98.73%
3	24	2	91.67%
6	54	0	100.00%
8	44	0	100.00%
10	20	1	95.00%
11	6	0	100.00%
13	5	0	100.00%
19	25	2	92.00%
20	5	0	100.00%

Place nr.	Total decisions	Failed decisions	Performance
1	79	0	100.00%
3	24	0	100.00%
4	23	1	95.65%
9	54	0	100.00%
11	44	0	100.00%
13	20	1	95.00%
14	6	0	100.00%
17	5	0	100.00%
22	25	2	92.00%
24	5	0	100.00%

RESULTS

To evaluate the performance of a localization method, we had to analyze the relationship between the predicted location (based upon the ECG algorithm) and the real location (based upon ablation site) of the AcP.

Table 1 present this relation for both algorithms, indicating the sensitivity and specificity for all AcP locations. Ablation sites are represented in vertical and the predicted locations in horizontal direction.

All AcP location predictions were analyzed in order to determine the strong and weak decision points of both localization methods. The result of this analysis is deductible from Table 1. During performance determination the contribution of each decision point was taken into consideration in when it was possible. In some cases a failed decision made all further evaluation meaningless. Such cases are indicated having no result. The last column shows the decision performance for all possible algorithmic branches. Table 2 represents the global accuracy of each decision point.

DISCUSSION

A penetrative study of the localization method demands much more than a simple evaluation of its overall performance, which proves to depend mostly on the least efficient decision points.

A deeper investigation of Table 2 reflects that applying Arruda localization method the estimation errors often (about 80% of the cases) appear due to wrong decision at places 3, 10, and 19 (see Figure 3). All of these decision points contain an aVF sign test. Furthermore, it is observable that in the diagram none of the aVF sign tests perform reliably.

The main goal of the modified localization method was to eliminate the most sensible decision points and thus to increase the overall performance. Unfortunately, this elimination proves to be easy only theoretically, due to the followings:

The depolarization wave's sphere-symmetric propagation is highly deformed by the fact that the right ventricle wall (4-5 mm) is much thinner than the left (10-12 mm) or the septal (7-8 mm) one. Taking into consideration the fact that the depolarization wave propagates at 300-800 mm/sec, and the sign of the delta wave is investigated over 20-millisecond intervals, it is obvious that vectorial calculations cannot accurately describe the behavior of the right ventricle.

The three leads used by Arruda, namely I, V1, and aVF, are almost orthogonal. Although V2 should perform better than V1, no significant difference proved achievable. Unfortunately, we have to admit that despite Arruda's decision points 10 and 19, which detect the septal and right regions, are sensible, they cannot be wisely replaced within the scope of the 12 lead ECG. This is illustrated also in Figure 3, where the aVF tests are encircled to indicate the necessity of further development.

The thicker wall of the left ventricle allows us to apply vectorial analysis; therefore, in this region significant improvements are possible.

Arruda's leads I and aVF (decision points 1 and 3) do not assure visibility in the direction of x axis. The modified method (decision points 1, 3, and 4) uses the z direction instead, which according to Table 2 at least partially improves the performance. Decision point 6 needs further refinement, because its 95% accuracy is relatively low.

These considerations were tested on our database, and must be used with care, due to the followings:

- Our database was too small to guarantee a solid statistical confirmation;
- The selected patients were not 100% representative (due to the small number);
- We used only few recordings from one patient, so the WPW syndrome could manifest in other way (it could change its behavior hourly).

CONCLUSION

Finally, we can conclude the followings:

- Arruda's leads used for the detection of septal and right regions seem to give optimal solution in the scope of 12-lead ECG.
- Significant improvements are achieved in the left ventricle.
- A more accurate analysis requires further studies and a significantly larger database.

The about 95% recognition rate we considered quite good, but a straightforward comparison with other localization methods (Boersma, Moran, Mont, & Brugada, 2002) performed on the same database would be beneficial in order to create a better noninvasive localization method .

A high localization performance of noninvasive methods is relevant because they can enlighten the necessary invasive interventions and also reduce the discomfort caused to the patient.

ACKNOWLEDGMENT

The research has been supported by the Hungarian National Research Fund, Grants OTKA T029830, OTKA T042990, Domus Hungarica Scientiarium et Artium, and Sapientia KPI.

REFERENCES

Arruda, M., McClelland, J., Wang, X., Beckman. K., et al. (1998). Development and validation of an ECG algorithm for identifying accessory pathway ablation site in Wolff-Parkinson-White syndrome. *J Cardiovascular Electrophysiology, 9*, 2-12.

Boersma, L., Moran, E. G., Mont, L., & Brugada, J. (2002). Accessory pathway localization by QRS polarity in children with Wolff-Parkinson-White syndrome. *J Cardiovasc Electrophysiol, 13*(12), 1222-1226.

Fitzpatrick, A. P., Gonzales, R. G., Lesh, M. D., et al. (1994). New algorithm for the localization of accessory atrioventricular connections using a baseline electrocardiogram, *J Am Coll Cardiol, 23,* 107-116.

Guanglin, L., & Bin, H. (2001). Localization of the site of origin of cardiac activation by means of a heart-model-based electrocardiographic imaging approach. *IEEE Trans. Biomed. Eng., 48,* 660-669.

Guize, L., Soria, R., Chaouat, J. C., et al. (1985). Prevalence and course of Wolff-Parkinson-White syndrome in population of 138,048 subjects. *Ann Med Interne Paris*, 136474-136489.

Huiskamp, G., & Greensite, F. (1997). A new method for myocardial activation imaging. *IEEE Trans Biomed Eng., 44*, 433–44446.

Reddy, G. V., & Schamroth, L. (1987). The localization of bypass tracts in the Wolff-Parkinson-White syndrome from the surface electrocardiogram. *Am Heart J, 113,* 984.

Rosenbaum, F. F., Hecht, H. H., Wilson, F. N., & Johnston, F. D. (1945). The potential variation of the thorax and the esophagus in anomalous antrioventricular excitation (Wolff–Parkinson–White syndrome). *Am Heart J, 29,* 281-326.

Shahidi, A.V., Savard, P., & Nadeau, R. (1994). Forward and inverse problems of electrocardiography: Modeling and recovery of epicardial potentials in humans. *IEEE Trans. Biomed. Eng., 41,* 249-256.

Szilágyi, S. M., Benyó, Z., & Frigy, A. (2004). Sensibility analysis of the ARRUDA localization method. *Periodica Politechnica Univ Timisoara, Transactions on Automatic Control and Computer Science*, *49(63)*/2, 129-132.

Wellens, H. J. J., Fare, J., & Bar, F. W. (1987). The Wolff-Parkinson-White syndrome. In W. J. Mandel (Ed.), *Cardiac arrhythmias. Their mechanisms, diagnosis and management* (pp. 274-296). Philadelphia, PA: JP Lippincott.

Wellens, H. J. J., Brugada, P., Penn, O. C., et al. (1990). Pre-excitation syndromes: Clinical presentation, course and therapy. In D. P. Zipes & J. Jalife (Eds.), *Cardiac electrophysiology. From cell to bedside* (pp. 691-702). Philadelphia, PA: WB Saunders Co.

Wolff, L., Parkinson, J., & White, P. (1930). Bundle branch block with short PR interval in healthy young people prone to paroxysmal tachycardia. *Am Heart J, 5,* 685-704.

Yee, R., Klein, G. J., & Guiraudon, G. M. (1995). The Wolff-Parkinson-White syndrome. In D. P. Zipes & J. Jalife (Eds.), *Cardiac electrophysiology: From cell to bedside* (pp. 1199-1214). Philadelphia, PA: WB Saunders Co.

KEY TERMS

Accessory Pathway: An abnormal electric conduction route that connects the atria with the ventricles.

Atria: The chambers of the heart that collect the venous blood.

Electrocardiogram (ECG): A graphic produced by an electrocardiograph, which records the electrical activity of the heart over time.

QRS Complex: A structure on the ECG signal that corresponds to the depolarization of the ventricles.

Ventricles: The pumping chambers of the heart.

Wolff-Parkinson-White (WPW) Syndrome: A syndrome of pre-excitation of the ventricles of the heart due to an accessory pathway that causes an abnormal electrical communication from the atria to the ventricles.

M

Modified Beamspace Method for the Spatial Filtering of Magnetoencephalographic Data

Tolga Esat Özkurt
University of Pittsburgh, USA

Mingui Sun
University of Pittsburgh, USA

Robert J. Sclabassi
University of Pittsburgh, USA

INTRODUCTION

Magnetoencephalography (MEG) and electroencephalography (EEG) are noninvasive measurement tools that provide high temporal resolution on the units of milliseconds to investigate neuronal activity in the brain. MEG and EEG measure the magnetic fields and the scalp electrical potentials produced by the current sources, respectively. MEG systems utilize gradiometer and/or magnetometer connected to superconducting quantum interference devices (SQUIDs) as sensors, while EEG systems are metal electrodes connected to differential amplifiers. It is well-known that the main advantage of MEG over EEG is its invulnerability to the distortions caused by various layers of cerebrum such as skull, scalp, muscle, and cerebrospinal fluid (Baillet, Mosher, & Leahy, 2001; Hamalainen, Hari, Ilomoniemi, Knuutila, & Lounasmaa, 1993). Estimating the neuronal source parameters mainly as source locations and magnitudes has been a great interest of researchers, with the aim of imaging the brain with a fine temporal resolution. The techniques for the extraction of these parameters are known as inverse methods. Various inverse methods, such as least-square estimation (Baillet et al., 2001), vector beamformers (Sekihara, Nagarajan, Poeppel, Marantz, & Miyashita, 2001), Multiple Signal Classification (MUSIC) (Mosher, Lewis, & Leahy, 1992), minimum norm solutions (Hauk, 2004), maximum likelihood-based solutions have been proposed in literature. Even in a noiseless environment, there is not a unique solution for the source parameters, which makes the problem ill-posed. Hence, all inverse methods depend on some assumptions about the sources, and when these assumptions are not appropriate in the unknown sources, satisfactory results may not be obtained.

Most information contained in MEG measurements reflects signals originated from the cortex because of the relatively short distances between the cortex and the magnetometers/gradiometers in the MEG system. Nevertheless, there have been some attempts to detect the sources arising from deep areas of the brain, such as brainstem (Parkkonen & Mäkelä, 2002), thalamus (Tesche, 1998), hippocampus (Tesche & Karhu, 2000), and cerebellum (Tesche & Karhu, 1997). However, the success of these attempts was limited.

Before using any inverse method, preprocessing of data may dramatically increase the overall performance in terms of both accuracy and computational efficiency. Preprocessing methods are realized for different goals, including artifact removal, dimension reduction, and noise removal. Among these methods signal space projection (SSP) (Tesche, Uusitalo, Ilmoniemi, Huotilainen, Kajola, & Salonen, 1995; Uusitalo & Ilmoniemi, 1997) relies on removing contributions from undesired sources by exploiting their spatial structures. Because of the assumed orthogonality of source and artifact spaces, the interesting part is also mostly modified while suppressing the undesired part. Popular and efficient signal processing methods such as independent component analysis (James & Gibson, 2003; Vigario, Sarela, Jousmaki, Hamalainen, & Oja, 2000) and adaptive filtering (Ahmar & Simon, 2005; Constantin, Richard, Lengelle, & Soufflet, 2005) have also been employed to cope with similar problems. These techniques mostly suffer from assumptions of strong uncorrelatedness or independency. Gross and Ioannides (1999) defined, evaluated and compared various linear transformation techniques, including the beamspace approach. The beamspace method relies on projecting data by maximizing the power with an orthogonal transformation matrix that spans the space of the leadfield obtained by forward modeling computations. This preprocessing technique has recently been shown to increase the performance of main source localization algorithms (Rodriguez, Baryshnikov, Van Veen, & Wakai, 2006). Signal space separation (SSS), provided first by Taulu and Kajola (2005), is also a novel technique designed principally for removing interferences from MEG measurements. Since there is no charge on the sensor array volume, Laplace's equation becomes satisfied for magnetic scalar potential. Utilizing this fundamental law of physics, the SSS method decomposes the recorded magnetic field into two parts using vector spherical harmonic basis functions: one for the signals coming from the inside the sensor array volume, and the other coming from the outside it. The method can effectively remove the external interferences without imposing unrealistic assumptions by estimating the coefficients in the least-linear square sense.

In this article, we propose a novel preprocessing method in the spherical harmonics domain to decompose the MEG signal into specific parts, whose sources arise from user-prescribed concentric spherical regions. A particular case of this approach is presented to separate the data into parts corresponding to deep and superficial regions of the brain, without using inverse solutions. Throughout this article, plain italics denote scalars, lower case boldface symbols denote vectors, uppercase boldface symbols denote matrices, superscripts T and H stand for transpose and Hermitian transpose, and $\|.\|$, $tr(.)$, * indicate Euclidean norm, and trace and complex conjugate operations, respectively.

BACKGROUND

Our algorithm to decompose the signal into deep and superficial regions is highly dependent on the SSS and beamspace methods, which were briefly introduced in the previous section. In this section, we supply the necessary definitions to support the algorithm development.

The mapping from current sources to the magnetic fields in a noiseless environment can be described as:

$$b_k(t) = \int_{\Omega} \mathbf{h}_k(\mathbf{r}') \cdot \mathbf{j}(\mathbf{r}')\, d\Omega \qquad (1)$$

where Ω is the whole source space, $\mathbf{r}'$ stands for the source locations, and $\mathbf{h}_k$ is the leadfield mapping of the current sources $\mathbf{j}$ to the magnetic field measurements at the k^{th} sensor, denoted as $b_k(t)$. The SSS method decomposes an M channel MEG signal:

$$\mathbf{b} = [b_1(t), b_2(t), ..., b_M(t)]^T =$$

$$\sum_{l=0}^{\infty}\sum_{m=-l}^{l} \alpha_{lm}\mathbf{x_{lm}} + \sum_{l=0}^{\infty}\sum_{m=-l}^{l} \beta_{lm}\mathbf{y_{lm}} = \mathbf{b}_{in} + \mathbf{b}_{out} \qquad (2)$$

into two components, where inner component $\mathbf{b}_{in}$ corresponds to source locations $\|\mathbf{r}'\| < R$, and the outer component $\mathbf{b}_{out}$ corresponds to $\|\mathbf{r}'\| > R$, and R is the radius of the sensor array. Based on the quasistatic approximation of Maxwell's equations, this separation is achieved using two different sums of vector spherical harmonic functions $\mathbf{x}_{\mathrm{lm}}$ and $\mathbf{y}_{\mathrm{lm}}$ (Taulu & Kajola, 2005). Spherical harmonic functions are orthonormal eigenfunctions of the Laplacian operator on the spherical surface, which makes them naturally useful tools for MEG signal processing, and hence has been suggested in literature for different purposes. For instance, Popov (2002) proposed a continuation of MEG data around the surface of the sensor array using a spherical harmonics expansion. They were also commonly utilized for the computation and approximation of the MEG forward problem (Jerbi, Mosher, Baillet, & Leahy, 2002; Nolte, Fieseler, & Curio, 2001; Nolte, 2003).

Equation 2 may be rewritten in an algebraic form as:

$$\mathbf{b} = \mathbf{S}\omega \qquad (3)$$

where the $(M \times p)$ dimensional basis functions matrix $\mathbf{S} = [\mathbf{S}_{in}\ \ \mathbf{S}_{out}]$ comprises inner and outer basis functions:

$$\mathbf{S}_{in} = [\mathbf{x}_{1,-1}, \mathbf{x}_{1,1}, \mathbf{x}_{2,-2}, ..., \mathbf{x}_{L_{in},L_{in}}] \qquad (4)$$

$$\mathbf{S}_{out} = [\mathbf{y}_{1,-1}, \mathbf{y}_{1,1}, \mathbf{y}_{2,-2}, ..., \mathbf{y}_{L_{out},L_{out}}]$$

and the $(p \times 1)$ coefficient vector $\boldsymbol{\omega} = [\boldsymbol{\alpha}\ \boldsymbol{\beta}]^T$ contains the SSS coefficients for inner and outer parts:

$$\boldsymbol{\alpha} = [\alpha_{1,-1}, \alpha_{1,1}, \alpha_{2,-2}, ..., \alpha_{L_{in}, L_{in}}]$$

$$\boldsymbol{\beta} = [\beta_{1,-1}, \beta_{1,1}, \beta_{2,-2}, ..., \beta_{L_{out}, L_{out}}], \tag{5}$$

and if the total number of coefficients is denoted by $p = p_{in} + p_{out}$, then the number of inner coefficients is $p_{in} = L_{in}^2 + 2L_{in}$, and the number of outer coefficients is $p_{out} = L_{out}^2 + 2L_{out}$. It should be noted that the coefficients with $l=0$ and $m=0$ are dismissed in the expansion, since they represent the magnetic monopoles. Then, the SSS coefficients α for the inner part, and β for the outer part, may be estimated using a pseudo-inverse operation as $\hat{\boldsymbol{\omega}} = \mathbf{S}^{+}\mathbf{b}$. Then, the reconstruction of the signal that corresponds to inner sources is obtained easily by $\hat{\mathbf{b}}_{in} = \mathbf{S}_{in}\hat{\boldsymbol{\alpha}}$.

Beamspace method has typically been used for dimension reduction and noise elimination of MEG data (Rodriguez et al., 2006), before it is being processed to extract the source magnitudes and locations by an inverse method. It looks for a linear transformation matrix **T** that maximizes the power in a region of interest, and is obtained by a criterion:

$$\max_{\mathbf{T}} \ tr(\mathbf{T}^T\mathbf{G}\mathbf{T}) \ \text{subject to} \ \mathbf{T}^T\mathbf{T} = \mathbf{I} \tag{6}$$

where **G** is the so-called Gram matrix (Gross & Ioannides, 1999), having the second-order relation between the leadfields:

$$\mathbf{G} = \int_{\Omega} \mathbf{H}(\mathbf{r}')\mathbf{H}(\mathbf{r}')^T \, d\Omega \tag{7}$$

where $\mathbf{H}(\mathbf{r}') = [\mathbf{h}_1(\mathbf{r}'), \mathbf{h}_2(\mathbf{r}'), \ldots, \mathbf{h}_M(\mathbf{r}')]^T$ is the ($M \times 3$) leadfield matrix for all M sensors. Solution of Equation 6 produces a **T** that is obtained as a ($N \times M$) dimensional matrix, whose columns are the eigenvectors of the Gram matrix **G**, those corresponding to the greatest N eigenvalues ($N \leq M$). Hence, the transformed and dimension reduced signal $\tilde{\mathbf{b}}$ is obtained as $\tilde{\mathbf{b}} = \mathbf{T}^{\mathbf{T}}\mathbf{b}$. Rodriguez et al. (2006) provide the details of this criterion that allows for an optimal linear transformation of the magnetic field data in mean-square sense.

MODIFIED BEAMSPACE IN THE SSS DOMAIN

This section presents the development of our algorithm to constrain the MEG signal into parts corresponding to deep and superficial current sources. The schematic representation given in Figure 1 summarizes the main steps of our algorithm.

After the inner SSS coefficients are estimated from the MEG signal, our goal is to optimally manipulate them, such that the reconstruction of the manipulated coefficients will give rise to two components: deep component $\mathbf{b}_{\text{deep}}$, where the electrical current sources come from the spherical region $\|\mathbf{r}'\| < \hat{r}$, and a superficial component $\mathbf{b}_{\text{sup}}$, where the sources arise from the remaining region of the head as $\|\mathbf{r}'\| > \hat{r}$. Here, $\hat{r}$ denotes the separating radius, and is determined by the user such that $0 < \hat{r} < R$. The geometry of this separation is exhibited in Figure 2. This task cannot be achieved only by the SSS algorithm itself, since the condition of the SSS algorithm is a boundary that does not contain current sources like the sensor array volume. However, this is not the case for a region inside the head, since it will always contain neural sources.

In the spherical harmonics domain, the leadfield-like relation between the SSS coefficients α_{lm} and the inner current sources $\mathbf{j}_{in}(\mathbf{r}')$ is formulated (Taulu & Kajola, 2005) as:

$$\alpha_{lm} = \int_{\Omega_{in}} \boldsymbol{\lambda}_{lm}(\mathbf{r}') \cdot \mathbf{j}_{in}(\mathbf{r}') \, d\Omega_{in} \tag{8}$$

Here, λ_{lm}s are leadfield-like vectors directly related to the vector spherical harmonic function $\mathbf{x}_{lm}(\theta,\varphi)$ defined as (Taulu & Kajola, 2005):

$$\boldsymbol{\lambda}_{\mathbf{lm}}(\mathbf{r}') = \frac{i}{2l+1}\sqrt{\frac{l}{l+1}}\, r^l \mathbf{x}_{\mathbf{lm}}^*(\theta,\varphi)$$

$$\mathbf{x}_{\mathbf{lm}}(\theta,\varphi) = \frac{-1}{\sqrt{l(l+1)}}\left[\frac{mY_{lm}(\theta,\varphi)}{\sin\theta}\mathbf{e}_\theta + i\frac{\partial Y_{lm}(\theta,\varphi)}{\partial\theta}\mathbf{e}_\varphi\right] \tag{9}$$

$Y_{lm}(\theta,\varphi)$ is known to be the normalized scalar spherical harmonic function (Arfken, 1985). Since MEG is sensible only to the tangential components of the sources, $\mathbf{x}_{lm}(\theta,\varphi)$ does not have a radial component. Hence, the dimensionality is naturally reduced from three to two. Additionally, the vector spherical harmonics yield an orthogonal representation unlike the nonorthogonal representation by leadfield vectors of forward modeling.

Figure 1.

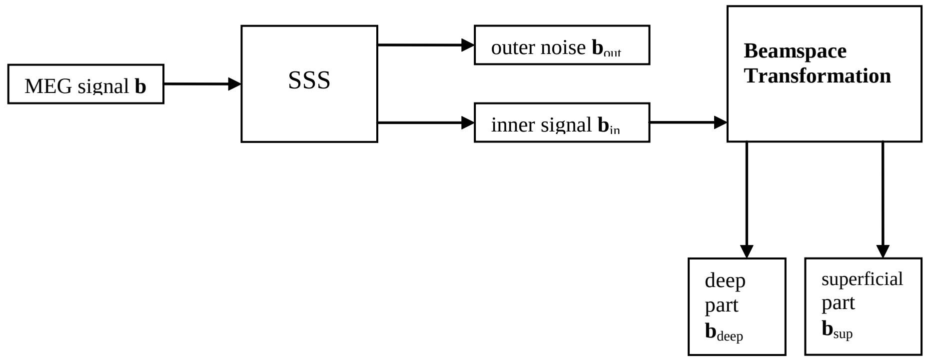
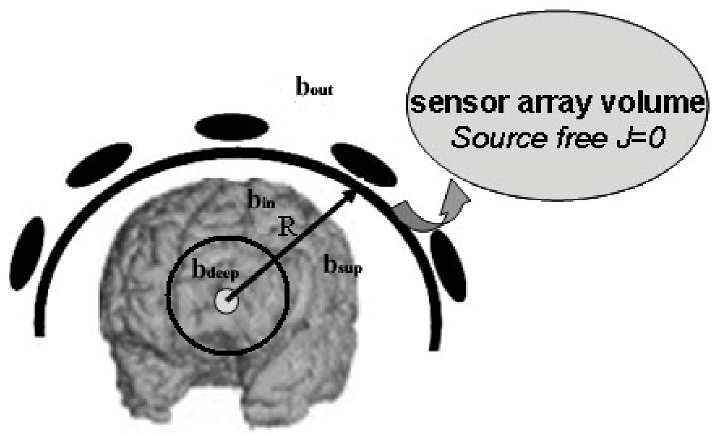


Figure 2.



If we assign the (l,m) indexed SSS coefficient as $a=1,\ldots,p$, and the (L,M) indexed SSS coefficient as $b=1,\ldots,p$ for simplicity, then the Gram matrix for the spherical harmonics domain can be defined as:

$$\left(\mathbf{G}\right)_{ab}^{\;S} = \int_{\Omega} \boldsymbol{\lambda}_{lm}(\mathbf{r'}) \cdot \boldsymbol{\lambda}_{LM}(\mathbf{r'})\, d\Omega = \int_{\Omega} \boldsymbol{\lambda}_{a}(\mathbf{r'}) \cdot \boldsymbol{\lambda}_{b}(\mathbf{r'})\, d\Omega$$

$$\mathbf{G}^{S} = \int_{\Omega} \boldsymbol{\Lambda}(\mathbf{r'})\boldsymbol{\Lambda}(\mathbf{r'})^{H}\, d\Omega \tag{10}$$

where $\boldsymbol{\Lambda}(\mathbf{r'}) = [\boldsymbol{\lambda}_{\mathbf{1\text{-}1}}(\mathbf{r'}), \boldsymbol{\lambda}_{\mathbf{11}}(\mathbf{r'}), \boldsymbol{\lambda}_{\mathbf{2\text{-}2}}(\mathbf{r'})\ldots, \boldsymbol{\lambda}_{\mathbf{LL}}(\mathbf{r'})]^{T}$ has dimension $(p_{in} \times 2)$. In practice, p_{in} is chosen to be less than M to estimate the SSS coefficients properly (i.e., after the signal is filtered by SSS, its dimension is already reduced from M to p_{in}).

We apply a beamspace criterion to transform the inner coefficients α, so that the power for the deep component is maximized, and for the superficial component, it is minimized:

$$\max_{\mathbf{T}} \frac{tr(\mathbf{T}^{T}\mathbf{G}_{d}\mathbf{T})/v_{d}}{tr(\mathbf{T}^{T}\mathbf{G}_{s}\mathbf{T})/v_{s}} \quad \text{with the constraint } \mathbf{T}^{T}\mathbf{T} = \mathbf{I} \tag{11}$$

Here, $\mathbf{G}_{d}$ and $\mathbf{G}_{s}$ are Gram matrices, and v_{d} and v_{s} are volumes for deep and superficial components, respectively. Because of the orthonormality property of the vector spherical harmonics (Arfken, 1985), the Gram matrices are diagonal. Then, the a^{th} rows and b^{th} columns of them are computed as follows:

$$(\mathbf{G}_d)_{ab} = \iint_{\varphi\,\theta} \int_{r=0}^{\hat{r}} \left(\frac{i}{2l+1}\sqrt{\frac{l}{l+1}} r^l \mathbf{x}_{lm} * (\theta,\varphi) \right) \bullet \left(\frac{-i}{2L+1}\sqrt{\frac{L}{L+1}} r^L \mathbf{x}_{LM} (\theta,\varphi) \right) r^2 \sin\,\theta dr d\theta d\varphi \qquad (12)$$

$$(\mathbf{G}_s)_{ab} = \iint_{\varphi\,\theta} \int_{r=\hat{r}}^{R} \left(\frac{i}{2l+1}\sqrt{\frac{l}{l+1}} r^l \mathbf{x}_{lm} * (\theta,\varphi) \right) \bullet \left(\frac{-i}{2L+1}\sqrt{\frac{L}{L+1}} r^L \mathbf{x}_{LM} (\theta,\varphi) \right) r^2 \sin\,\theta dr d\theta d\varphi \qquad (13)$$

Equation 12 reduces to the matrix form (Ozkurt, Sun, & Sclabassi, 2006).

$$(\mathbf{G}_d)_{ab} = \delta_{lL}\delta_{mM} \frac{1}{(2l+1)^2}\frac{l}{l+1} \iint_{\varphi\,\theta} |\mathbf{x}_{lm}(\theta,\varphi)|^2 \sin\,\theta dr d\theta d\varphi \int_{r=0}^{\hat{r}} r^{2l+2} dr = \delta_{lL}\delta_{mM} \frac{1}{(2l+1)^2}\frac{l}{l+1}\frac{\hat{r}^{2l+3}}{2l+3} = \delta_{ab}\, y(l) \qquad (14)$$

where $y(l) = \frac{1}{(2l+1)^2}\frac{l}{l+1}\frac{\hat{r}^{2l+3}}{2l+3}$.

Similarly, Equation 13 shrinks to:

$$(\mathbf{G}_s)_{ab} = \frac{1}{(2l+1)^2}\frac{l}{l+1}\frac{R^{2l+3}-\hat{r}^{2l+3}}{2l+3}\delta_{ab} \qquad (15)$$

The solution of Equation 11 leads to a matrix **T** whose columns are eigenvectors of:

$$\mathbf{G}_f = \frac{v_s}{v_d}\mathbf{G}_s^{-1/2}\mathbf{G}_d(\mathbf{G}_s^{-1/2})^T \qquad (16)$$

corresponding to the largest eigenvalues (Gross & Ioannides, 1999). Hence, combining Equations 14, 15, and 16, we obtain the final diagonal Gram matrix:

$$(\mathbf{G}_f)_{ab} = \frac{v_s}{v_d}G_s^{-1/2}G_d(G_s^{-1/2})^T\delta_{ab} = \frac{v_s}{v_d}\frac{\hat{r}^{2l+3}}{R^{2l+3}-\hat{r}^{2l+3}}\delta_{ab} \qquad (17)$$

Then all eigenvectors of $\mathbf{G}_f$ will constitute an identity matrix. Accordingly, Equation 17 suggests a transformation that selects the first inner SSS coefficient $\boldsymbol{\alpha}_{deep} = [\alpha_{1,-1}, \alpha_{1,1}..., \alpha_{s,s}]$ where

$$\frac{v_s}{v_d}\frac{\hat{r}^{2l+3}}{R^{2l+3}-\hat{r}^{2l+3}}$$

is sufficiently large. All the remaining coefficients are eliminated. The reconstruction of α_{deep} will yield an estimate of a component $\hat{\mathbf{b}}_{deep}$ that corresponds to deep components. Similarly, the estimation of superficial sources is achieved by minimizing the reciprocal of the first term in Equation 11, and accordingly, beamspace transformation selects the last inner SSS coefficients where $\boldsymbol{\alpha}_{\sup} = [\alpha_{L_{in}-s+1,-(L_{in}-s+1)},..., \alpha_{L_{in},L_{in}}]$ for the reconstruction of the superficial component of the magnetic signal $\hat{\mathbf{b}}_{sup}$. When l is considered as spatial frequency, the beamspace result can be interpreted as "sharp" low-spatial frequency filtering for the deep part, and high spatial frequency filtering for the superficial part (see Figure 3a). This result agrees with Equation 9, since it shows that the greater the l values of the SSS coefficients are, the greater the contribution from the superficial sources. Although sharp spatial filtering results from the optimality in mean-square sense, weighting coefficients uniformly may not be so reliable, since Equation 9 clearly indicates that the contributions from the sources are rather exponential. To overcome this problem, we suggest an eigenvalue-weighted filtering of the selected SSS coefficients. Hence, the beamspace transformation matrix **T** is modified as **T'=TD,** where **D** denotes the diagonal matrix whose diagonal elements are the selected largest eigenvalues of $\mathbf{G}_f$. This modification leads to a simple manipulation of the SSS coefficients for deep sources as:

$$(\tilde{\alpha}_{lm})_{deep} = \alpha_{lm}\frac{v_s}{v_d}\frac{\hat{r}^{2l+3}}{R^{2l+3}-\hat{r}^{2l+3}} \qquad (18)$$

and reconstructing the signal of interest with the obtained modified coefficients for the deep part:

$$\hat{\mathbf{b}}_{deep} = \mathbf{S}_{in}\left(\tilde{\boldsymbol{\alpha}}_{deep}\right)^T \qquad (19)$$

Estimation of the superficial part is similarly realized by the modification given below:

$$(\tilde{\alpha}_{lm})_{sup} = \alpha_{lm} \frac{v_d}{v_s} \frac{R^{2l+3} - \hat{r}^{2l+3}}{\hat{r}^{2l+3}}$$
$$\hat{\mathbf{b}}_{sup} = \mathbf{S}_{in} \left(\tilde{\boldsymbol{\alpha}}_{sup}\right)^T . \qquad (20)$$

This transformation corresponds to a rather smoother spatial filtering as depicted in Figure 3b.

Some MEG data was simulated in order to verify the accuracy and the performance of the proposed algorithm. Its sensor characteristic was chosen to be compatible with the Elekta Neuromag® 306 channel system, which measures the magnetic signal using 204 planar gradiometers and 102 magnetometers at a sampling frequency of 1 KHz. A spherical head model of radius $R=9$ cm was utilized in the simulations. Two sources dipole were created at coordinates (3,1,2) cm for the deep source, and (6,5,4) cm for the superficial source. Duration of the field data was 600 ms, which corresponds to a data length of 600 samples. The waveforms are sinusoids with frequencies 2 Hz for the deep source, and 10 Hz for the superficial source (Figure 4a). Since there are no external interferences for the synthesized data, we determined $L_{out}=0$, and choosing $L_{in}=10$ was enough to represent the input signal with the SSS coefficients. Therefore, the dimension of the 306 channel magnetic measurement is reduced to $p_{in}=120$. Note that in this simulation, the energies of deep sources and superficial sources contributing to total data were same.

We applied our algorithm to the simulated data with the properties described above for a separating radius $\hat{r}$ =3.6 cm, which specifies a spherical boundary between the deep source ($r_d \approx 3.32$cm), and for the superficial source ($r_s \approx 8.77$cm) of the head model (Figure 4 b). Figure 5 provides overlapping plots of all channels for the simulated data (a), estimated deep part (b), and superficial part (c) with the modified inner SSS coefficients as described in Equations 19 and 20 using the proposed eigenvalue weighted beamspace approach. It can be observed that the algorithm extracted deep and superficial sources reasonably well, although there is still a noticeable superficial leakage in the estimated deep component. In Figure 6, we provide two channels for a better visual comparison. In Figure 6a, it is observed that our algorithm achieves the extraction of the superficial contribution from a channel that is close to deep sources. In Figure 6b, we exhibit how the algorithm resolves the deep component from a channel that has a strong superficial source contribution.

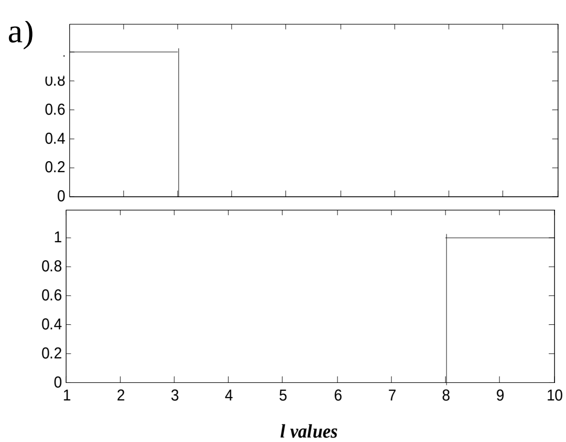
FUTURE TRENDS

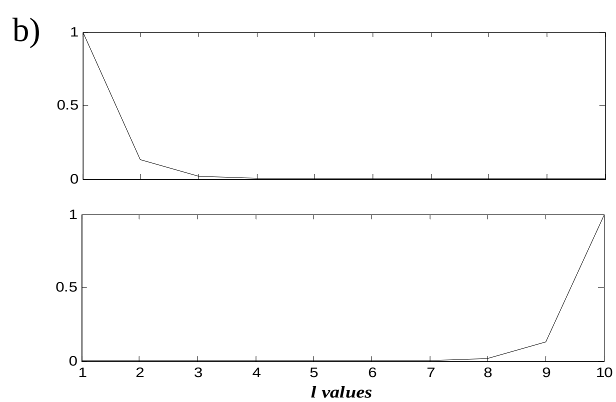
Deeper sources, such as the ones arising from the basal ganglia, are related to information processing and storage within the brain, and hence possess the potential importance to be investigated. However, they were not studied thoroughly by MEG community because

Figure 3.

Spatial filtering of the coefficients according to the **classical** beamspace criterion for deep (top panel) and superficial (bottom panel) sources.

Spatial filtering of the coefficients according to the **modified** beamspace criterion for deep (top panel) and superficial (bottom panel) sources.





of their weakness and longer distance to head, when compared with the cortex. Our study enables the raw MEG data to be decomposed, corresponding to sources within different radial concentric layers. This will help to reduce the masking effect of the strong cortical signals on much weaker signals arising from the deeper regions, and provide a better resolution for the region of interest. Various well-known inverse methods afterwards can be applied to decomposed data to localize the neural activities according to the user's interest. There have been other suggestions, such as in Kim, Kobayashi, and Uchikawa (2003) to constrain the MEG data sources to regions of interest with taking account of the spectral information of the sources. However, our algorithm exploits only the spatial properties, and does not need to assume or impose any prior information about the source waveforms.

CONCLUSION

Decomposing the MEG data to specific locations of the brain without using inverse solutions may provide invaluable insights about the properties of the data. The SSS algorithm is an attempt to realize it by obtaining coefficients for interesting sources inside the sensor array, and the external noise outside it. We propose to extend this algorithm to constrain the data to some desired locations inside the brain. This study showed that with a simple manipulation of the inner SSS coefficients, the signal can be separated into parts that correspond to deep and superficial sources. This simplicity of the derived formulation comes from the natural appropriateness to spherical domain and orthogonality properties of the SSS basis functions that are directly related to the vector spherical harmonics. Beamspace method is used to obtain the optimality in mean-square sense for the modification of the SSS coefficients. It should be noted that it would not be easy to do these computations with classical lead field functions (i.e., since they are not orthogonal, one would have to first determine necessary coordinates by dividing the deep (or superficial) source volume into grids and compute the Gram matrix discretely, unlike the proposed method). Moreover, with our approach, while the beamspace transformation matrix is being obtained, one does not have to deal with the sensor configurations and dimension reduction (i.e., choosing the eigenvectors of the Gram matrix that correspond to the largest eigenvalues). All these procedures are already handled by the SSS method.

This algorithm may also be developed to decompose the signal into different regions of interest, such as left and right hemispheres, and bottom and top parts of the head. In this case, the vector harmonic basis will no longer be orthogonal when integrated across these source regions, which could make it much more difficult to obtain the Gram matrix. We are currently investigating efficient and optimal solutions using spherical harmonics to constrain the data into arbitrarily determined locations. Additionally, parameters such as the separating radius and different energy ratios of deep and superficial parts will affect the performance of the algorithm. We will report the results of these considerations in the near future.

Figure 4.

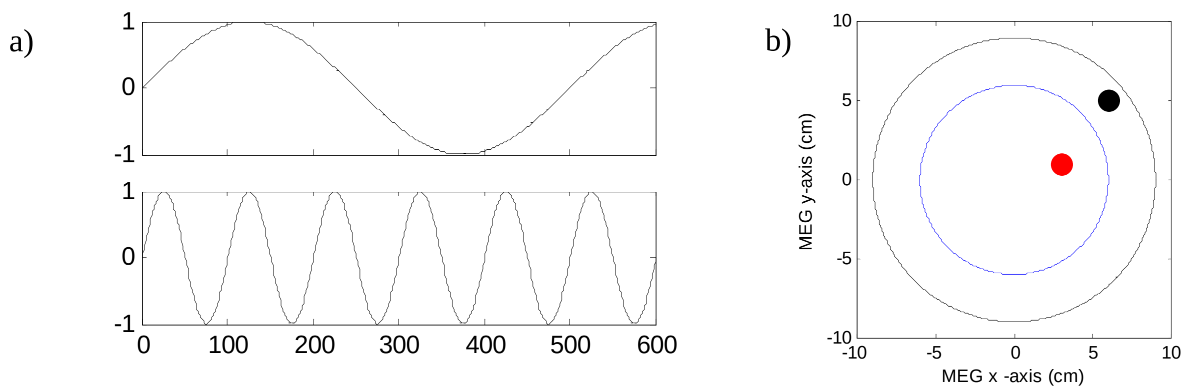
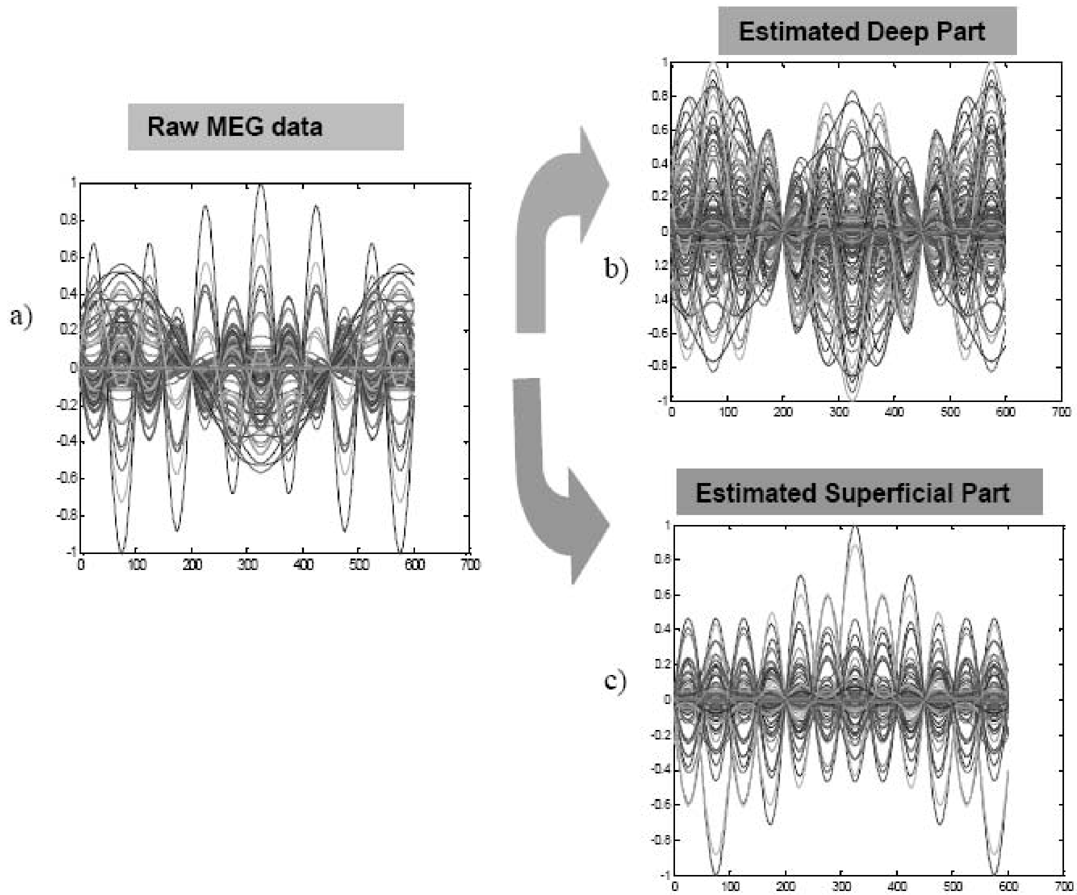


Figure 5.



ACKNOWLEDGMENT

This work was supported in part by National Institutes of Health Grants EB002309, and NS/MH38494 and Computational Diagnostics Inc. The authors would like to acknowledge the Center for Advanced Brain Magnetic Source Imaging (CABMSI) of the University of Pittsburgh Medical Center (UPMC), and Elekta Neuromag® for their support to conduct the MEG tests. They also thank Wenyan Jia, Eliezer Kanal, and Gusphyl Justin for their helpful suggestions and comments.

REFERENCES

Ahmar, N. E., & Simon, J. Z. (2005). MEG adaptive noise suppression using fast LMS. In *Proceedings of the 2nd IEEE EMBS Conference on Neural Engineering* (pp. 5–8).

Arfken, G. (1985). *Mathematical methods for physicists*. New York: Academic Press.

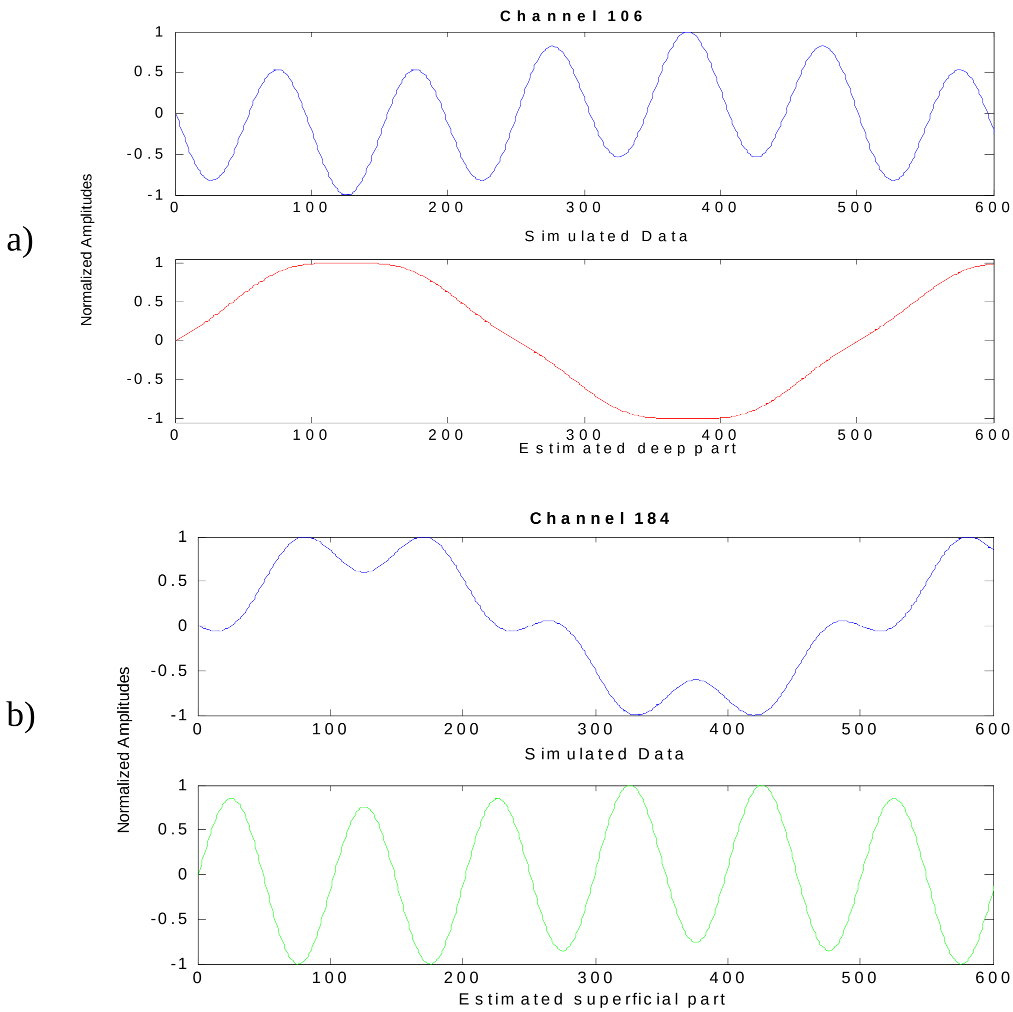
Baillet, S., Mosher, J. C., & Leahy, R. M. (2001). Electromagnetic brain mapping. *IEEE Signal Processing Magazine, 18(6),* 14–30.

Constantin, I., Richard, C., Lengelle, R., & Soufflet, L. (2005). Regularized kernel-based wiener filtering, application to magnetoencephalographic signals denoing. In *Proceedings of IEEE International Conference of Acoustics, Speech, Signal Processing* (pp. 289–292).

Gross, J., & Ioannides, A.A. (1999). Linear transformations of data space in MEG. *Physics in Medicine and Biology, 44,* 2081-2097.

Hamalainen, M., Hari, R., Ilomoniemi, R. J., Knuutila, J., & Lounasmaa, O. V. (1993). Magnetoencephalography—Theory, instrumentation and applications to noninvasive studies of the working human brain. *Reviews of Modern Physics, 65*(2), 413–497.

Figure 6.



Hauk, O. (2004). Keep it simple: A case for using classical minimum norm estimation in the analysis of EEG and MEG data. *Neuroimage, 21,* 1612–1621.

James, C. J., & Gibson, O. J. (2003). Temporally constrained ICA: An application to artifact rejection in electromagnetic brain signal analysis. *IEEE Transactions on Biomedical Engineering, 50*(9), 1108–1116.

Jerbi, K., Mosher, J. C., Baillet, S., & Leahy, R. M. (2002). On MEG forward problem using multipolar expansions. *Physics in Medicine and Biology, 47,* 523–555.

Kim, B. S., Kobayashi, K., & Uchikawa, Y. (2003). Seperation of overlapping activity in first and second somatosensory evoked fields with 3-D MEG measurement. *IEEE Transactions on Magnetics, 39*(5), 3387–3389.

Mosher, J. C., Lewis, P. S., & Leahy, R. M. (1992). Multiple dipole modeling and localization from spatio-temporal MEG data. *IEEE Transactions on Biomedical Engineering, 39*(6), 541–557.

Nolte, G., Fieseler, T., & Curio, G. (2001). Perturbative analytical solutions of the magnetic forward problem for realistic volume conductors. *Journal of Applied Physics. 89*(4), 2360–2369.

Nolte, G. (2003). The magnetic lead field theorem in the quasi-static approximation and its use for MEG forward problem. *Physics in Medicine and Biology, 49*(22), 3637-3652.

Özkurt, T. E., Sun, M., & Sclabassi, R. J. (2006). Beamspace magnetoencephalographic signal decomposition in spherical harmonics domain. In *Proceedings of the*

28th International Conference of the IEEE Engineering in Medicine and Biology Society (pp. 5743–5746).

Parkkonen, L., & Mäkelä, J. (2002). MEG sees deep sources: Measuring and modeling brainstem auditory evoked fields. In *Proceedings of the 13th International Conference on Biomagnetism* (pp. 107–109).

Popov, M. (2002). Data continuation for the explicit solution of an inverse biomagnetic problem. *IEEE Transactions on Magnetics, 38*(6), 3620-3632.

Rodriguez, A., Baryshnikov, B. V., Van Veen, B. V., & Wakai, R. T. (2006). MEG and EEG source localization in beamspace. *IEEE Transactions on Biomedical Engineering, 53*(3), 431–441.

Sekihara, K., Nagarajan, S. S., Poeppel, D., Marantz, A., & Miyashita, Y. (2001). Reconstructing spatio-temporal activities of neural sources using an MEG vector beamformer technique. *IEEE Transactions on Biomedical Engineering, 48*(7), 76–771.

Taulu, S., & Kajola, M. (2005). Presentation of electromagnetic multichannel data: The signal space separation method. *Journal of Applied Physics, 97*(12), 124905.1–124905.10.

Tesche, C.D. (1998). Detecting activity from deep brain areas with magnetoencephalographic arrays. In *Proceedings of the International Conference of IEEE Engineering In Medicine and Biology Society* (Vol. 20, No. 4, pp. 2201–2204).

Tesche, C.D., & Karhu, J. (1997). Somatosensory evoked magnetic fields arising from sources in the human cerebellum. *Brain Research, 744*(1), 23–31.

Tesche, C.D., & Karhu, J. (2000). Theta oscillations index human hippocampal activation during a working memory task. In *Proceedings of the National Academic of Sciences of the United States of America* (Vol. 97, No. 2, pp. 919-924).

Tesche, C.D., Uusitalo, M. A., Ilmoniemi, R. J., Huotilainen, M., Kajola, M., & Salonen, O. (1995). Signal-space projections of MEG data characterize both distributed and well-localized neuronal sources. *Electroencephalography and Clinical Neurophysiology, 95*, 189–200.

Uusitalo, M. A., & Ilmoniemi, R. J. (1997). Signal space projection method for separating MEG or EEG into components. *Medical & Biological Engineering & Computing, 35,* 135–140.

Vigario, R., Sarela, J., Jousmaki, V., Hamalainen, M., & Oja, E. (2000). Independent component approach to the analysis of EEG and MEG recordings*.* *IEEE Transactions on Biomedical Engineering, 47*(5), 589–593.

KEY TERMS

Beamspace: It is a preprocessing method to reduce the dimension of the MEG signal by using a linear transformation matrix that is obtained using leadfields.

Forward Problem: It corresponds to calculating the magnetic field outside the head from a given primary source distribution within the brain.

Gram Matrix: It describes the deterministic second order relations among leadfields for different source locations.

Inverse Methods: These methods are utilized to estimate the source locations and strengths from EEG / MEG signals.

Leadfield: It denotes a vector mapping from a unit current source to magnetic measurement on a sensor located in a particular position.

Magnetoencephalography (MEG): It measures extracranial magnetic fields induced by electrical currents in the brain, noninvasively. The very weak magnetic fields of the brain are sensed by superconducting quantum interference devices (SQUIDs).

Signal Space Separation (SSS): A preprocessing method that decomposes the MEG signal into inner source signal and outer noise by using spherical harmonic basis functions.

Spherical Harmonics: They are orthonormal eigenfunctions of the Laplacian operator on the spherical surface, and useful tools to represent EEG/MEG signals on the head for different cases.

M

Monitoring and Controlling of Healthcare Information Systems (HIS)

Stefan M. Graeber
Saarland University, Germany

Ansgar Kutscha
Diakonie Hospital Schwaebisch Hall gGmbH, Germany

BACKGROUND

Information management (IM) at a health care institution encompasses the management of information, the management of application systems, and the management of information and communication technology whether computer supported or not, that is, IM provides function, performance, and quality of HIS. Management means, as well, the responsible persons and organizational units as the tasks of planning, directing, and monitoring HIS. IM has to be done systematically to enable an orderly processing of information coherent with the goals of the health care institution.

While planning and directing are supported comprehensively by basic methods of strategic planning and project management (Brigl, Ammenwerth, Dujat et al., 2005; Haux, Winter, Ammenwerth, & Brigl, 2004; Winter, Ammenwerth, Bott et al., 2001), the monitoring is neglected sometimes and thus insufficiently supported (Ammenwerth, Ehlers, Hirsch, & Gratl, 2007). As nevertheless a continuous and careful monitoring is a very important task in interaction with all other management tasks, we will define the relevant terms and describe the most significant concepts and methods.

Monitoring

Generally, *monitoring* of HIS means the continuous observation of whether the directives and objectives defined in the strategic information management plan will be reached, and whether the HIS is able to fulfill the required tasks. Therefore, the IM must be able at any time to assess the state of the HIS using quality criteria which can be derived from the objectives. Its results affect directing and planning again by feedback mechanisms.

The tasks of monitoring may be linked to the strategic level (auditing HIS quality as defined by means of strategic information management plan's directives and goals as well as quality of the strategic management process itself), the tactical level (check whether the initiated projects are running as planned and whether they will produce the expected results), and the operational level (verifying the proper working and effectiveness of all HIS components) (Haux et al., 2004, p. 182-184).

Nowadays the management tasks providing an excellent service for all users of HIS are embraced by the term *IT service management* (ITSM). There are several frameworks describing an architecture for installing and maintaining ITSM. The most known framework is the *IT infrastructure library* (ITIL) (www.itil.org). It is a set of best practices enabling organizations to deliver their services more efficiently and thus at last to reach for a maximum of customer (patient) satisfaction. ITIL may be regarded as a guideline for monitoring of HIS.

IT Controlling

Management decisions require information or data. In this context, the part of IM delivering information needed as basis for management decisions is called *IT controlling*. For this purpose, IT controlling applies different approaches and methods, for example, the continuous measuring and interpretation of indicators and characteristic values explaining the current state of HIS, or the realization of evaluation studies. Thus, among other IT controlling, comprises following tasks (the terms reference model, indicator, and evaluation project are outlined below):

- Defining and operationalizing objectives (in cooperation with partners of IM)

- Defining models, selection and application of reference models
- Defining indicators and appropriate values
- Planning, initiating, and continuous measurement of indicators
- Planning and performance of evaluation projects
- Reporting results
- Analyzing results (which may influence all preceding steps by feedback-mechanism) (in cooperation with partners of IM)
- Preparing decisions (in cooperation with partners of IM)

Depending on tasks and questions, different methods of information acquisition are applied. Sometimes one performs ad-hoc-studies (field studies) to find hypotheses, to get some insight in the features of performance measuring, or to detect problems and deficiencies (screening). An example of screening is a survey performed to discover the problems with a new nursing documentation system two weeks after installation. More important are the continuous data collection via indicators and occasional deeper investigations (evaluation).

Indicators and Characteristic Values

Indicators are variables whose values (characteristics values) represent an aspect of HIS. To discern good and bad quality of information processing and to assess the achievement of goals, one has to compare the current value of indicator with one or more predefined reference values. With standardized indicators comparisons between different HIS become possible. Relevant aspects may be all components of HIS, for example, strategy, projects, quality, processes, functionality, or parts of IT infrastructure. The indicators can be qualitative (e.g., user satisfaction), quantitative non-monetary (e.g., failure time), or quantitative monetary (e.g., cost). They should be specified as follows:

- Comprehensive description, including its purpose and correlation with the objectives
- Data source, measurement procedure, and algorithms (when indicators are derived from others)
- Characteristic values and reference values (limits), for example, corresponding to quality goals
- Time interval for measurement (e.g., daily, weekly)
- Time of availability
- Responsible organizational unit and person(s)
- Procedure to check adequateness, completeness, and correctness of results afterwards (evaluation of indicators)

To describe complex aspects of HIS, several indicators have to be considered. The combination of logically associated indicators is called an *indicator system*. The Balanced Scorecard (see next paragraph) is an example for an indicator system.

Balanced Scorecard

A suitable tool for management and controlling is the *Balanced Scorecard* (BSC) (Kaplan & Norton, 1992, 2000). At first developed for strategic management of an organization, the BSC can be applied for purposes of monitoring too. It provides feedback around both the internal business processes and external outcomes in order to continuously improve strategic performance and results. When fully deployed, the balanced scorecard transforms strategic planning into the "nerve center" of an enterprise. Using BSC the manager may establish a "balanced" situation between the traditional financial measures and other success factors of an organization. The BSC combines the continuous measurement of performance with a reviewing and refinement strategy as well an ongoing evaluation process.

The basic terms are so-called *perspectives*, for example, potentials (learning and growth), internal processes, customers, and financial perspective. Depending on the view, the term customer can vary: customers from the view of health care organization are mostly patients, customers from the view of IM are users of IT applications. For each perspective, strategic objectives, indicators, and measures to achieve these objectives must be defined.

Models and Reference Models

A *model* is a simplified representation of a section of the real world (subject area). Models are developed for understanding, analysis, or improvement of subject area. Depending on purposes of modeling the relevant aspects of subject area are selected and built in only. Models become an important tool for the IM, because they help overwhelm the complexity of HIS.

A metamodel can be considered as a toolbox for building models. With general metamodels like ARIS (www.ids-scheer.com/international/english/products/53961), one can describe processes independent of the business field. The three-layer graph-based metamodel (3LGM²) is a specialized tool for static HIS modeling. The domain layer consists of functions and entity types. The logical tool layer focuses on application components, and the physical tool layer describes physical data processing components. Additionally several inter-layer-relationships are defined (Winter, Brigl, & Wendt, 2003). The 3LGM² allows building of models of HIS presenting their current state with adequate levels of granularity. So one can find out, for example, which functions are supported by which application systems, or the kind and scope of communication processes between different applications. In this manner, deficiencies may be detected, too, for example, gaps in information transfer. Additionally the 3LGM² allows to integrate indicator systems and to compute indicators along the relations between the model objects (Kutscha, Brigl, & Winter, 2006).

A *reference model* is a template for a well-defined class of models, that is, from a reference model concrete models are derived by modification, restraints, or additions (Haux et al., 2004, p. 73). Furthermore one can check by comparison, whether concrete models have the same features as the reference model, for example, completeness of functions. Therefore the use of a reference model can deliver relevant indicators. Depending on subject area, there are reference models for organizations, information systems, software, or procedures.

Reference models support the quality assessment better than models. They function like standards, that is, if the quality of reference model is proved, the quality of a concrete model may be made evident by comparison using quality indicators. The reference model itself should be evaluated also to provide that the model is a sufficient map of reality.

For the monitoring of HIS, one needs reference models for information systems to find out whether the business processes are supported by efficient tools, or the necessary information flows are established (Winter et al., 1999). But also reference models for procedures are useful (see examples provided).

Evaluation

Evaluation is the act of measuring or exploring properties of a health care information system (in planning, development, implementation, or operation), the result of which informs a decision to be made concerning that system in a specific context (Ammenwerth, Brender, Nykanen, Prokosch, Rigby, & Talmon, 2004). With the information acquired by evaluation, one can learn something, answer relevant questions, or make decisions. In this manner the screening mentioned above can be regarded as a "simple" evaluation. Evaluation of health care systems is a main topic of research in medical informatics (Brender, 2006; Friedman & Wyatt, 1997). Therefore, here some methodical principles are outlined only.

There is a wide spectrum of evaluation methods. It reaches from simple surveys to exactly planned longitudinal investigations. Some studies are performed like clinical trials. It means that one should try to apply principles of Good Clinical Practice (http://www.ich.org/LOB/media/MEDIA482.pdf). But often some techniques are not applicable (e.g., randomizing or blinding). Nevertheless reliable studies can be performed, if some recommendations are taken into account (Ammenwerth, Graeber, Herrmann, Buerkle, & Koenig, 2003). All decisions and steps should be documented in a detailed study protocol. An adequate study design and appropriate methods to answer the study questions have to be selected. The combination of quantitative and qualitative methods may be helpful. Wherever possible, validated evaluation instruments should be used. Often a multi-methodic and/or multidisciplinary approach is necessary (Ammenwerth et al., 2003).

Evaluation studies can be formative or summative. Formative evaluation strives to improve the HIS component under evaluation by providing the developers (and implementers) with feedback. Summative evaluation tries to demonstrate the outcome of a "mature" HIS component in routine (Friedman & Wyatt, 1997, p. 304).

Organization of IM

Organizational structures for IM differ considerably among health care institutions. In general, each institution should have an adequate organization for strategic, tactical, and operational IM, depending on its size, internal structure, and needs. Reference models for the organization of IM in hospitals are given in Haux et al. (2004, p. 187ff).

The relation between IM and IT controlling may be formed very differently. Between the two "poles" (IT controlling totally embedded without own directives vs. IT controlling as specific organizational unit beside IM) several structures are possible.

Next to organization the performance of IM plays an important role. For example, the performance level of IM can be assessed by the Capability Maturity Model (www.sei.cmu.edu/cmmi/). The IM should have a high maturity level, that is, it should function as *business enabler* for the hospital.

EXAMPLES

Table 1 shows perspectives (without financial perspective) and corresponding objectives from a BSC established at a German university hospital. The main perspectives of BSC were modified slightly. A part of the indicator tree for the subperspective *information management* and the appropriate objective *efficient support of processes by IT* is presented in Table 2. The indicator tree defines the indicators and their dependencies, which have to be measured for a perspective. Overall this BSC contains 11 composite performance indicators reflecting particular strategic objectives. These measures are expressed in terms of actual performance as percentage of expected performance. For further examples of BSC application, see Kaplan and Norton (2000).

Ammenwerth et al. (2007) developed a so-called HIS-monitor to describe the strengths and weaknesses of information processing at hospitals. This monitor is a matrix whose rows represent several quality indicators while the columns define the necessary processes and subprocesses of patient care. The matrix is filled out during standardized interviews with hospital staff. Thus, one can reveal the processes that are not sufficiently supported by IT.

ITIL is an example for a reference model for procedures. It defines the necessary management tasks at tactical level (service delivery, Figure 1) and operational level (service support, Figure 2). The advantage of ITIL for the alignment of IT with business objectives is reported (Kashanchi & Toland, 2006).

An example of a reference model for a hospital information system is the requirements index for information processing in hospitals (Ammenwerth, Buchauer, & Haux, 2002). German experts established the index in a consensus-based, top-down, and cyclic manner. Each functional requirement was derived from information processing functions and subfunctions of a hospital. It contains 233 functional requirements and 102 function-independent requirements. The functional requirements are structured according to the primary

Table 1. Perspectives and corresponding objectives for a BSC (examples)

Perspective	Objectives
Potentials Employees Infrastructure Organization *Information management* …	*Efficient support of processes by IT* Establishing interdisciplinary functional units …
Processes Medicine Research Teaching Administration …	Establishing structured and transparent workflows for patient care (e.g., clinical pathways) Better and faster implementation of current medical evidence …
Customers Patients and relatives Cooperating practitioners Cooperating hospitals Students Consultants …	Increasing the integration of in- and out-patient care Enhancement of user satisfaction …

M

Table 2. Part of the indicator tree

Function level

- Availability of application infrastructure
  - Failure time
    - Applications
    - Basic services
    - Hardware
    - Network
- Reaction time/problem-solving time
  - Number of problems
  - Number of problems solved within one day
- Completeness of medical record
  - Ratio of digital images
  - Ratio of released documents
- User satisfaction
  - Functionality
  - Availability
  - Support
- …

Dissemination level

- …

…

Figure 1. Tasks and objects of service delivery (from Ammenwerth et al., 2007)

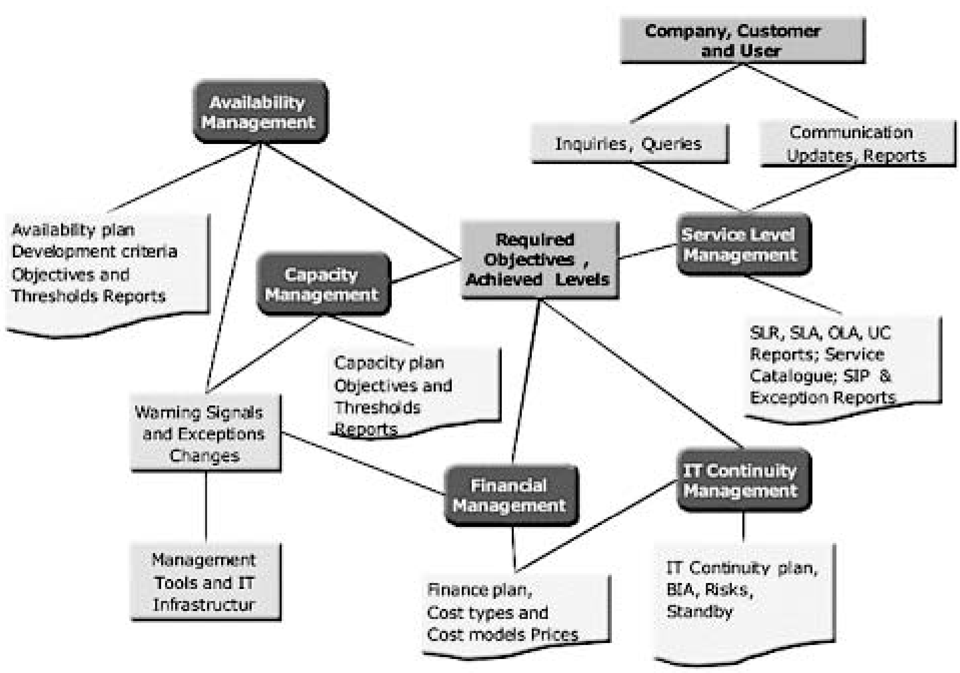
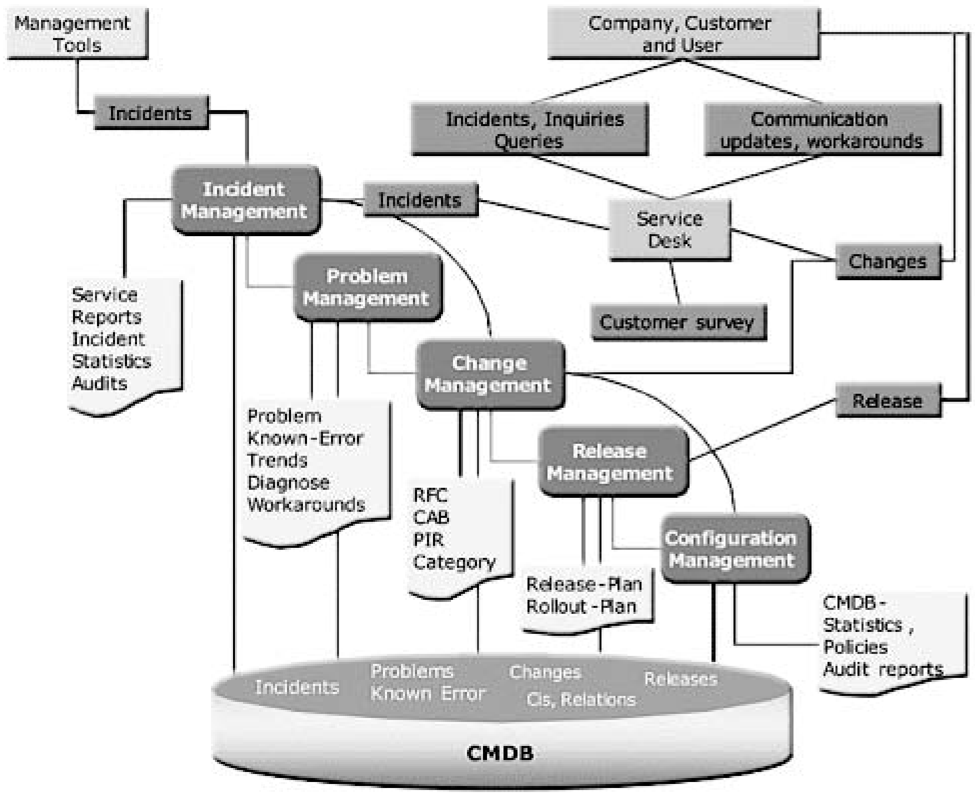


Figure 2. Tasks and objects of service support (from Ammenwerth et al., 2007)



Legend: CMDB = Configuration Management Data Base

care process from admission to discharge, handling of patient records, work organization and resource planning, hospital management, research, and education. For example, with the index, one can find out the ratio of implemented functions.

Only three examples of evaluation studies are mentioned here, two of those were conducted and published by the first author. A method for the selection of software for a subsystem of HIS is described in Graeber (2001). An investigation of user satisfaction with clinical workstations was published in Graeber (1997). Machan, Ammenwerth, and Schabetsberger (2006) published the rather comprehensive and understandable description of an evaluation project to assess the electronic transmission of medical findings. Evaluation projects of the last 20 years can be found in the Web-based inventory of evaluation studies in medical informatics (evaldb.umit.at). The database can be searched with keywords. Many entries may be used as patterns for successful evaluation studies in fields of HIS.

Problems and Challenges

Often an insufficient organization of IM is the main obstacle for an effective HIS monitoring. This problem includes the lack of highly qualified staff for ambitious management tasks. Sometimes the objectives of IM are not adequate defined or they are not consequently derived from the enterprise goals (Brigl et al., 2005). Also a general deficiency of awareness of the importance of monitoring and controlling as well as insufficient consideration of the needs of customers and users can inhibit the establishment of an IM resp. ITSM structure at health care institutions.

A further problem is the development from *hospital* information system to *health care* information system. A modern HIS encompasses many different health care organizations and institutions with different IT infrastructures. This complexity makes a comprehensive and effective IM very difficult.

Although the successful use of BSC in health care sector is described (Stewart & Bestor, 2000; Protti, 2002), its application for HIS monitoring yields some new aspects. Especially the following success factors have to be considered:

- The method must be strongly supported by the enterprise management.
- The application of BSC requires a lean management. Otherwise a schism may occur between the management and the need to get detailed understanding of work processes and motivation from staff (Brender, 2006, p. 86).
- An early and continuous communication with employees about the purpose and results of BSC is needed.
- All involved people must have an open mind for definition of perspectives, objectives, indicators, and measures as well for the interpretation of results.
- One should prefer less but significant indicators.
- IT-support of BSC is not compulsory, but helpful. One should strive for the integration of BSC in a data warehouse system.
- The BSC has to be evaluated periodically.

Although the effort for the development of BSC can be considerable, its use for monitoring is very functional. It helps to translate the vision of a health care institution and to gain consensus, to communicate the objectives, to link goals with strategy, to allocate resources, and to provide feedback and ensure learning. It is an efficient way to link day-to-day operating activities to the strategic objectives.

A suitable model of HIS is an important prerequisite for an efficient IM. With the 3LGM² not only technical and semantic aspects but also computer-based and paper-based information processing are integrated in the model. Deficiencies in the current state of HIS can be detected and hence the quality of information processing can be assessed. Such a model can be the basis of the strategic information management plan (Brigl et al., 2005) as well as of economic analyzes (Kutscha et al., 2006). A disadvantage of 3LGM² is the static view. For modeling of dynamic aspects a business process metamodel (e.g., ARIS at www.ids-scheer.com/international/english/products/53961) has to be used.

Contrarily to other branches, reference models for the information management in health care are seldom. A cause may be that most models are too specific and not suited as reference. These restrictions concern the BSC too. Therefore the development of valid reference models for HIS is a major challenge.

Rapid changes in technology (e.g., telehealth) and increasing complexity of HIS cause also new conditions for evaluation projects, for example.

- Studies must be carried out in different types and size sites, with different HIS components and different groups of users.
- People, organizational, social, and ethical issues must be taken into account.
- Evaluation aspects must be incorporated into all phases of a project.
- Importance and effect of formative evaluations are increasing.

Problems and challenges of evaluation projects in medical informatics are discussed detailed in Kaplan and Shaw (2004).

REFERENCES

Ammenwerth, E., Brender, J., Nykänen, P., Prokosch, H.U., Rigby, M., & Talmon, J. (2004). Visions and strategies to improve evaluation of health information systems—Reflections and lessons based on the HIS-EVAL workshop in Innsbruck. *Int J Med Inform., 73*, 479-491.

Ammenwerth, E., Buchauer, A., & Haux, R. (2002). A requirements index for information processing in hospitals. *Methods Inf Med., 41*(4), 282-288.

Ammenwerth, E., Ehlers, F., Hirsch, B., & Gratl, G. (2007). HIS-monitor: An approach to assess the quality of information processing in hospitals. *Int J Med Inform., 76,* 216-225.

Ammenwerth, E., Graeber, S., Herrmann, G., Buerkle, T., & Koenig, J. (2003). Evaluation of health information systems – problems and challenges. *Int J Med Inform., 71*(2-3), 125-135.

Brender, J. (2006). *Handbook of evaluation methods for health informatics*. Burlington: Elsevier Academic Press.

Brigl, B., Ammenwerth, E., Dujat, C., et al. (2005). Preparing strategic information management plans for hospitals: A practical guideline. *Int J Med Inform., 74,* 51-65.

Friedman, C.P., & Wyatt, J.C. (1997). *Evaluation methods in medical informatics*. New York: Springer.

Graeber, S. (1997). Application of clinical workstations: Functionality and usability. *Clin Perform Qual Health Care, 5*(2), 71-75.

Graeber, S.M. (2001). How to select a clinical information system. In *Proc AMIA Symp*. (pp. 219-223).

Haux, R., Winter, A., Ammenwerth, E., & Brigl, B. (2004). *Strategic information management in hospitals.* New York: Springer.

Kaplan, N., & Norton, D. (1992). The balanced scorecard – measures that drive performance. *Harv Bus Rev., 70*(1), 71-79.

Kaplan, N., & Norton, D. (2000). *The strategy-focused organization: How balanced scorecard companies thrive in the new business environment*. New York: McGraw-Hill.

Kaplan, B., & Shaw, N.T. (2004). Future directions in evaluation research: People, organizational, and social issues. *Methods Inf Med., 43*(3), 215-231.

Kashanchi, R., & Toland, J. (2006). Can ITIL contribute to IT/business alignment? *Wirtschaftsinformatik, 48*(5), 340-348.

Kutscha, A., Brigl, B., & Winter, A. (2006). Modeling economic aspects of hospital information systems to give decision support for strategic information management. *Stud Health Technol Inform., 124,* 672-678.

Machan, C., Ammenwerth, E., & Schabetsberger, T. (2006). Evaluation of the electronic transmission of medical findings from hospitals to practitioners by triangulation. *Methods Inf Med., 45*(2), 225-233.

Protti, D. (2002). A proposal to use a balanced scorecard to evaluate information for health: An information strategy for the modern NHS (1998-2005). *Comput Biol Med., 32*(3), 221-236.

Stewart, L.J., & Bestor, W.E. (2000). Applying a balanced scorecard to health care organizations. *Journal of Corporate Accounting & Finance, 11*(3), 75-82.

Winter, A., Ammenwerth, E., Bott, O.J., et al. (2001). Strategic information management plan: The basis for systematic information management in ospitals. *Int J Med Inform., 64,* 99-109.

Winter, A., Brigl, B., & Wendt, T. (2003). Modeling hospital information systems. Part 1: The revised three-layer graph-based meta bodel 3LGM2. *Methods Inf Med., 42*(5), 544-551.

Winter, A.F., Winter, A., Becker, K., Bott, O., Brigl, B., Gräber, S. et al. (1999). Referenzmodelle für die Unterstützung des Managements von Krankenhausinformationssystemen. *Informatik, Biometrie und Epidemiologie, 30*(4), 173-189.

Retrieved February 12, 2008, from www.itil.org.

Retrieved February 12, 2008, from www.ids-scheer.com/international/english/products/53961.

ICH Harmonised Tripartite Guideline—Guideline for good clinical practice E6. (2006). Retrieved February 12, 2008, from http://www.ich.org/LOB/media/MEDIA482.pdf

Retrieved February 12, 2008, from www.sei.cmu.edu/cmmi/

Ammenwerth, E., & de Keizer, N. (2006). *A Web-based inventory of evaluation studies in medical informatics.* Retrieved February 12, 2008, from evaldb.umit.at

KEY TERMS

Balanced Scorecard: The balanced scorecard (BSC) is a *management* tool that enables organizations to clarify their vision and strategy and translate them into action. Using BSC, the manager may establish a "balanced" situation between the traditional financial measures and other success factors of an organization. The BSC combines the continuous measurement of performance with a reviewing and refinement strategy as well an ongoing evaluation process. The basic terms are so-called perspectives, for example, potentials (learning and growth), internal processes, customers, and financial perspective. For each perspective, strategic objectives, indicators, and measures to achieve these objectives must be defined.

Evaluation: Evaluation is the act of measuring or exploring properties of HIS (in planning, development, implementation, or operation), the result of which informs a decision to be made concerning that system in a specific context. Contrarily to continuous data collection, evaluation studies are carried out to answer special questions, usually in form of a project with a clear time limit. When possible, for such studies the same criteria and methods as for clinical trials should be applied.

Indicators and Characteristic Values: Indicators are variables whose values (characteristics values) represent an aspect of HIS. To discern good and bad quality of information processing and to assess the achievement of goals, one has to compare the current value of indicator with one or more predefined reference values.

IT Controlling: In this context, the term controlling (or more precisely IT controlling) is restricted to the tasks of information acquisition and data collection as basis for decisions of IM. For this purpose IT controlling applies different approaches and methods, for example, the continuous measuring and interpretation of indicators and characteristic values explaining the current state of HIS, or the realization of evaluation studies.

IT Service Management: Beside the internal business processes, the focus of monitoring should be the external outcome, as well. To provide an excellent service for all users of HIS and at last to reach a maximum of customer (patient) satisfaction, nowadays the appropriate management tasks are embraced by the term IT service management (ITSM). There are several frameworks describing an architecture for installing and maintaining ITSM. The most known framework is the IT infrastructure library (ITIL), which defines the necessary management processes at tactical level (service delivery) and operational level (service support).

Monitoring: Monitoring of HIS means the observation as to whether the directives and objectives defined in the strategic information management plan will be achieved, and whether the HIS is able to fulfill the required tasks. Therefore the IM must be able at any time to assess the state of the HIS using quality criteria which can be derived from the objectives. The tasks of monitoring may be linked to strategic level (monitoring of the achievement of strategic information management plan), tactical level (monitoring of projects), and operational level (operational monitoring, that is, verifying proper working and effectiveness of all HIS components).

Reference Model: A reference model presents a kind of model patterns for a certain class of aspects. It can be used to derive a specific model or for purpose of comparison. Comparing a specific model of HIS (or subsystem) with a reference model congruencies and differences may be stated and used for planning and direction of HIS.

MRI Induced Heating on Pacemaker Leads

Eugenio Mattei
Istituto Superiore di Sanità, Italy

Giovanni Calcagnini
Istituto Superiore di Sanità, Italy

Michele Triventi
Istituto Superiore di Sanità, Italy

Federica Censi
Istituto Superiore di Sanità, Italy

Wolfgang Kainz
Center for Devices and Radiological Health, Food and Drug Administration, USA

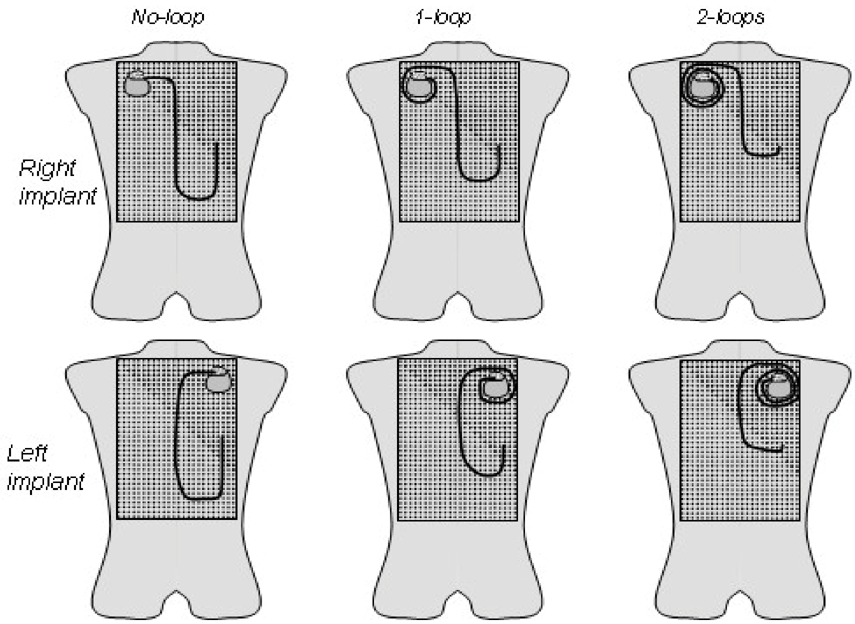
Howard Bassen
Center for Devices and Radiological Health, Food and Drug Administration, USA

Pietro Bartolini
Istituto Superiore di Sanità, Italy

INTRODUCTION

Magnetic resonance imaging (MRI) is a widely accepted tool for the diagnosis of a variety of disease states. The presence of a metallic implant, such as a cardiac pacemaker (PM), or the use of conductive structures in interventional therapy, such as guide wires or catheters, are currently considered a strong contraindication to MRI (Kanal, Borgstede, Barkovich, Bell, Bradley, Etheridge, Felmlee, Froelich, Hayden, Kaminski, Lester, Scoumis, Zaremba, & Zinninger, 2002; Niehaus & Tebbenjohanns, 2001; Shellock & Crues, 2002). Potential effects of MRI on PMs' implantable cardioverter defibrillator (ICDs) include: force and torque effects on the PM (Luechinger, Duru, Scheidegger, Boesiger, & Candinas, 2001; Shellock, Tkach, Ruggieri, & Masaryk, 2003); undefined reed-switch state within the static magnetic field (Luechinger, Duru, Zeijlemaker, Scheidegger, Boesiger, & Candinas, 2002); potential risk of heart stimulation and inappropriate pacing (Erlebacher, Cahill, Pannizzo, & Knowles, 1986; Hayes, Holmes, & Gray, 1987); and heating effects at the lead tip (Achenbach, Moshage, Diem, Bieberle, Schibgilla, & Bachmann, 1997; Luechinger, Zeijlemaker, Pedersen, Mortensen, Falk, Duru, Candinas, & Boesiger, 2005; Sommer, Vahlhaus, Lauck, von Smekal, Reinke, Hofer, Block, Traber, Schneider, Gieseke, Jung, & Schild, 2000). In particular, most of the publications dealing with novel MRI techniques on patients with implanted linear conductive structures (Atalar, Kraitchman, Carkhuff, Lesho, Ocali, Solaiyappan, Guttman, & Charles, 1998; Baker, Tkach, Nyenhuis, Phillips, Shellock, Gonzalez-Martinez, & Rezai, 2004; Nitz, Oppelt, Renz, Manke, Lenhart, & Link, 2001) point out that the presence of these structures may produce an increase in power deposition around the wire or the catheter. Unfortunately, this increased local specific absorption rate (SAR) is potentially harmful to the patient, due to an excessive temperature increase which can bring living tissues to necrosis. The most direct way to get a measure of the SAR deposition along the wire is by using a temperature probe: the use of fluoroptic® thermometry to measure temperature has become "state-of-the-art," and is the industry standard in this field (Shellock, 1992; Wickersheim et al., 1987). When the investigation involves small objects and large spatial temperature gradients, the measurement of the temperature increase and of the local SAR may become inaccurate, unless several precautions are taken. It seems obvious to: (1) evaluate the error associated with temperature increase and SAR measurements; (2) define a standard protocol for probe positioning, which minimizes the error associated with temperature measurement.

Figure 1. Sketches of the six implant configurations tested in a commercial MRI using the human-shaped phantom; no-loop, one-loop, and two-loop for both right and left pectoral implants



BACKGROUND

In most publications dealing with heating of conductive structures during MRI, the reported temperature increase is significantly different, even in cases where the experimental set-up is apparently similar. For example, Achenbach et al. (1997) reported a temperature increase of 63.1°C for a PM lead; Rezai, Finelli, Nyenhuis, Hrdlicka, Tkach, Sharan, Rugieri, Stypulkowski, and Shellock (2002) observed 25.3°C at the end of a deep brain stimulation electrode; Roguin, Zviman, Meininger, Rodrigues, Dickfeld, Bluemke, Lardo, Berger, Calkins, and Halperin (2004) reported a maximum increase of 5.7°C at 3.54 W/Kg whole body specific absorption rate (WB-SAR). Sommer et al. (2000), with a WB-SAR of 1.3 W/Kg, obtained temperature increases ranging from 0.1 to 23.5°C, depending on the electrode type. Several factors influence the degree of heating: (1) the WB-SAR has been shown to correlate to the temperature increase; (2) the cooling effect of the blood around the leads is generally not quantified; (3) the length and the geometric structure of the lead; and (4) the implant location.

In addition, since thin linear structures such as PM leads may generate temperature gradients which cannot be neglected, with respect to the physical dimension of temperature probes, also the relative positioning of the temperature probe to the lead tip significantly affects the measurement, and partially explains the inconsistency of the results in literature.

The aim of this article is first to identify the optimal positioning of fluoroptic® probes to measure the maximum heating of the tip of a PM lead. Then, we use these results to perform measurements in a real MRI system, in order to investigate the effect of the placement of the PM, and of the lead geometry.

MEASUREMENT OF TEMPERATURE AND SAR

We used two experimental setups: first, we performed temperature measurements on a PM lead tip inside a rectangular box phantom, using different types of contact positions between the probe and the lead tip. Second, we used a human-shaped phantom with a PM, and its leads exposed to a commercial MRI scanner.

PROBE POSITIONING

A PVC box (28 × 20 × 26 cm) was filled with a gelled saline (gel) solution. A grid was submerged in the phantom gel to support the implant and maintain consistent separation distances between the implant, phantom material surface, and temperature probes. The phantom material gel is composed of the following materials in percentages by weight: 2% hydroxy-ethyl-cellulose (*HEC*), 0.36% sodium chloride, and the rest, water. The gel has a conductivity of 0.59 Sm^{-1}, a permittivity of 79 at 64 MHz, and a heat capacity of 4178.3 J $Kg^{-1}K^{-1}$ (ASTM, 2004). SAR and temperature measurements were performed on the tip of a 62 cm-long monopolar lead (S80TM, *Sorin Biomedica CRM*, Italy). A radiofrequency (RF) signal was injected into the lead tip through a coaxial cable, connected to the lead. The outer conductor (signal ground) was connected to a 1x20x10 mm silver plate located on one side of the PVC box, so that the current flows in the gel from the lead tip to the plate. The lead was arranged in the phantom material 5 cm below the gel surface, simulating an implant in the human body. The distance between the silver plate and the lead tip was set at 7 cm. The temperature measurements were performed using a fluoroptic® thermometer (*Luxtron*, Model 3100, USA), with resolution of 0.1°C, operating at eight samples per second. Three sinusoidal excitations were applied: 25, 64, and 128 MHz, which nearly correspond to the RF field used in 0.5, 1.5, 3 T static field MRI systems. Signals were generated by a RF generator (*Rhode & Schwartz*—SMT 06), and then amplified (*RFPA*—RF 06100-6, France); a power meter (*Rhode & Schwartz NRT*—Z14) was connected to the output of the amplifier, in order to measure the average power and the reflection coefficient. Four configurations were studied (Figure 2):

- Transversal contact between the side of the probe and the circular surface of the lead tip (Figure 2, reference contact);
- Transversal contact between the tip of the probe and the side surface of the electrode (tip-to-side);
- Axial contact between the tip of the probe and the circular surface of the lead tip (tip-to-tip);
- Axial contact between the tip of the probe and the side surface of the electrode (side-to-side).

The effect of the position of the probes on the temperature measurements was expressed as the percentage error, in respect to the configuration giving the maximum temperature increase:

$$\Delta T\ \%\ error = \frac{\Delta T_{probe} - \Delta T_{max}}{\Delta T_{max}} \cdot 100 \qquad (1)$$

where ΔT_{probe} is the temperature increase measured by the probes in different configurations, and T_{max} the maximum of these values.

Local SAR was estimated as the slope (*dT/dt*) of the line that best fits the initial temperature increase. This estimation was assumed valid when the Pearson coefficient r^2 was greater than 0.98 (IEEE, 1991). The underestimation in SAR measurement, due to the different contact configurations as expressed in terms of percentage error, in respect to the maximum value of SAR:

$$SAR\ \%\ error = \frac{SAR_{probe} - SAR_{max}}{SAR_{max}} \cdot 100 \qquad (2)$$

where SAR_{probe} is the SAR measured by the probes in different configurations, and SAR_{max} the maximum of these values.

Our experiments showed how the position of the temperature probe significantly affects the SAR measure: transversal contact between the side of the temperature probe and the circular surface of the lead tip (Figure 2, reference contact) is the configuration which leads to the highest measured temperature. We define the temperature measured with other positions than the one leading to the maximum value as an underestimation (1). The highest temperature underestimation was obtained at 25 MHz, and it decreases as the frequency increases, regardless of the temperature probe contact. The configurations "tip-to-side" and "tip-to-tip" resulted in a temperature underestimation ranging from 28% to 39%. The underestimation associated with the side-to-side contact was significantly lower (4%–7%). These differences can be explained by the position of the temperature active element within the probe tip. Different temperature probe placement configurations result in different orientations of the temperature active element in relation to the PM lead tip. Therefore, the temperature active element can be exposed to different temperatures, especially in areas with high spatial temperature gradients.

The underestimation associated with the estimation of local SAR at the lead tip showed a similar behavior: transversal contact between the side of the temperature probe and the circular surface of the lead tip (Figure 2, reference contact) is the configuration measuring the highest SAR (1444 W/Kg). The underestimation associated with other temperature probe configurations is reported in Figure 2, upper panel. In the worst case, the SAR underestimation can be up to 75%. The effect of the frequency on SAR underestimation was similar to that observed for the temperature.

HUMAN-SHAPED PHANTOM

Experiments in real MRI scanning system were performed using a human-shaped torso simulator made at the Department of Technology and Health of the National Institute of Health in Rome. The simulator consists of a torso-shaped transparent PVC phantom of the size of a 70 kg male. Internal volume of the torso is 32 litre (ASTM, 2004). A PVC grid is mounted inside the torso to support the PM, the PM lead, and the temperature probes. The torso was filled with the same gel as described above.

The actual lead geometry and PM placement can vary from patient to patient. The PM can be located in the left or right pectoral region. Since the length of the lead may not fit the patient's anatomy and size, the exceeded length is usually wrapped near or around the PM. In the experiments on the human-shaped torso simulator, the geometry of the implant reproduced left and right PM placements. For each implant, three lead paths have been tested (Figure 1): without lead loop around the PM (no-loop configuration), and with the lead forming one or two loops around the PM (one-loop

Figure 2. SAR and temperature underestimation: Comparison among measurements made by a temperature probe in transversal contact with the PM lead tip (reference contact) and underestimation obtained with other contact configurations ("tip-to-side," "tip-to-tip," "side-to-side")

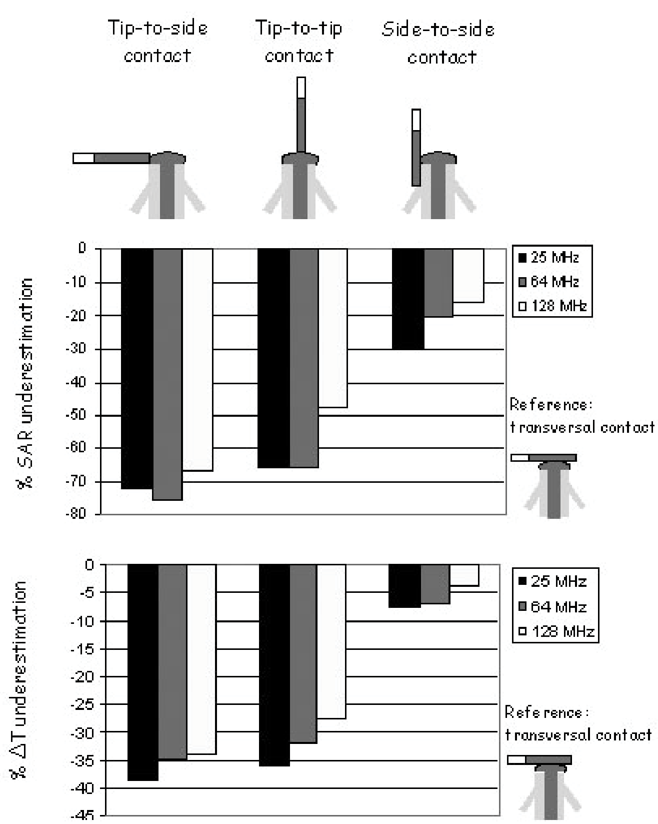


Table 1. Main parameters of the MRI clinical sequences used during the human-shaped phantom experiments

Siemens sequence name	Sequence type	TR (ms)	TE (ms)	Flip Angle	Length (s)
FLASH (short)	Spoiled gradient echo	960	40	20	53
FLASH (long)	Spoiled gradient echo	960	40	20	417
TrueFISP (short)	Steady state free process	3.78	1.89	54	38
TrueFISP (long)	Steady state free process	3.78	1.89	54	379
HASTE (short)	Single shot (turbo fast) spin echo	1190	83	150	42
HASTE (long)	Single shot (turbo fast) spin echo	1190	83	150	402

configuration and two-loop configuration, respectively). For each lead path, the length and the position of the linear section of the lead, as well as the lead tip position, were kept constant. Only the PM was moved from the left to the right pectoral location. Experiments were performed on a 1.5 T (static field) *Magnetom Sonata Maestro Class* scanner (*Siemens*, Germany). The main parameters of the sequences used are summarized in Table 1. The MRI parameters (TR, TE, and Flip Angle) were adjusted to reach a WB-SAR of 2 W/Kg, estimated by the scanner. The temperature increase and local SAR values for the experiments using a real MRI scanner are reported in Table 2. As for the clinical sequences, the gradient echo did not induce detectable temperature increase in all the configurations tested. Sequences with relatively high whole body SAR, led to a temperature

Table 2. Temperature increase (°C) and lead tip SAR (W/kg) of human-shaped phantom experiments; average whole body SAR calculated by the scanner is also reported

Siemens sequence name		Left pectoral implant			*SAR* (W/kg)*	Right pectoral implant			*SAR* (W/kg)*
		No loop	*1-loop*	*2-loops*		*No loop*	*1-loop*	*2-loops*	
Flash (short)	Temp.	0.10	<0.10	<0.10	0.02	<0.10	<0.10	<0.10	0.02
	SAR	(---)	(---)	(---)		(---)	(---)	(---)	
Flash (long)	Temp.	0.10	<0.10	<0.10	0.02	<0.10	<0.10	<0.10	0.02
	SAR	(---)	(---)	(---)		(---)	(---)	(---)	
Trufisp (short)	Temp.	4.20	0.65	0.40	1.72	8.17	1.93	0.62	1.70
	SAR	(1055)	(---)	(---)		(2192)	(---)	(---)	
Trufisp (long)	Temp.	6.17	0.97	0.60	1.70	12.26	2.50	0.96	1.70
	SAR	(1255)	(---)	(---)		(2375)	(536)	(---)	
Haste (short)	Temp.	6.92	1.03	0.72	1.96	9.41	2.09	0.74	1.94
	SAR	(2214)	(281)	(291)		(2871)	(643)	(365)	
Haste (long)	Temp.	6.35	0.92	0.68	1.70	11.91	2.68	1.01	1.72
	SAR	(1362)	(---)	(---)		(2345)	(641)	(---)	

**SAR: average whole body specific absorption rate computed by the scanner;*
(---) SAR not estimable, due to low temperature increase.

increase up to 12.3 °C. For left pectoral PMs, the lead area appears to be the major factor relevant for the heating. A high implant area (no loop) showed always a higher temperature increase than the one-loop and two-loop configurations.

Surprisingly, in right pectoral implants, the lead path seems to play the major role: no-loop configurations showed always a temperature increases significantly greater than the one- and two-loop configurations. In addition, the temperature increases observed were greater than those of the left implants. In both configurations (left and right), the temperature increase and local SAR were proportional to the whole body SAR reported by the scanner.

The comparison between short and long sequences showed that the major temperature increase occurred within the first minute. Experiments were repeated, changing MRI parameters such as the centre of view (chest, abdomen, and pelvis), and the field of view (200, 300, and 400 mm), without significant changes in the heating.

FUTURE INVESTIGATIONS

Our experimental measures showed the large number of variables which may be involved in the amount of heat generation and measurement of temperature increase. There are still some elements, such as the cooling effect of the blood around the leads, or the evaluation of the temperature and SAR errors yielded by other temperature probe types (e.g., surface and remote style probes), that need further investigations.

Moreover, this study investigated only the in-vitro heating of MRI on pacing leads. Other potential risks of MRI in PM recipients, such as fast pacing, inhibition of stimulation, or direct stimulation of the heart caused by the RF or gradient fields need further evaluation.

Owing to the large number of factors that influence the degree of heating, and that make it difficult to perform extensive and exhaustive experimental measures, the development of numerical models may represent a very useful approach to investigate the potential effects of MRI on PMs, ICDs, and other active implantable medical devices.

CONCLUSION

Fluoroptic® thermometry has become the most popular method to measure heating of metallic objects, due to MRI RF exposure, although some methodological issues have found limited attention so far. Among them, the positioning of the probes, as well as the estimation of the error in temperature and SAR measure, need to be investigated and standardized. Thin linear structures such as PM leads may generate temperature gradients which cannot be neglected, with respect to the physical dimension of the probes. As a consequence, specific methods and protocols are required to obtain accurate estimations of the maximum temperature and SAR in the region surrounding the lead tip.

Using temperature measurements on a physical model of a PM lead, we found that the positioning of temperature probes strongly affects temperature and SAR results. Due to the comparable dimension of the temperature probes with the PM lead tip, and due to the large spatial temperature gradient around the lead tip, temperature probes tend systematically to underestimate the real value of local temperature and SAR. In particular, using the SMM Luxtron fluoroptic® probes, we found that a transversal contact of the temperature probe with the lead tip always gives the highest temperature and SAR values. Using this configuration as reference, the underestimation yielded by other configurations may be as high as 39% for temperature, and 75% for SAR. Furthermore, we found less underestimation as the frequency increases. A deeper investigation of this issue was beyond the aim of this article.

The different contact configurations between fluoroptic® probes and the lead tip may explain, at least partially, the large variability of previous studies (Achenbach et al.,1997; Rezai et al., 2002; Roguin et al., 2004; Sommer et al., 2000).

Temperature and SAR measurements on the human-shaped phantom showed how the lead geometry affects significantly the amount of induced heating. In most cases, PMs are implanted in the left pectoral region. In this case, in the implant area (i.e. the area formed by the lead), the PM can, and a straight line connecting the PM tip to the PM can, plays a significant role on the lead tip heating. Implants with relatively large lead area (250-300 cm^2) exposed to a SAR of about 2 W/kg may experience temperature increase at the lead tip up to 12°C, and local SAR up to 3000 W/kg. Right pectoral implanted PMs cover smaller areas than left pectoral ones, but the temperature increases observed were higher, suggesting that the coupling to the electric field may also contribute to the heating. To explain such behaviour, resonance phenomena in various kinds of linear metallic leads and wires has been hypothesized by various groups (Duru, Luechinger, Scheidegger, Luscher, Boesiger, & Candinas, 2001; Nitz et al., 2001).

REFERENCES

Achenbach, S., Moshage, W., Diem, B., Bieberle, T., Schibgilla, V., & Bachmann, K. (1997). Effects of magnetic resonance imaging on cardiac pacemakers and electrodes. *Am. Heart J., 134*, 467–473.

American Society for Testing and Material (ASTM) Designation. (2004). *Standard test method for measurement of radio frequency induced heating near passive implants during magnetic resonance imaging.*

Atalar, E., Kraitchman, D.L., Carkhuff, B., Lesho, J., Ocali, O., Solaiyappan, M., Guttman, M.A., & Charles, H.K. Jr. (1998). Catheter-tracking FOV MR fluoroscopy. *Magn. Reson. Med., 40*, 865–872.

Baker, K.B., Tkach, J.A., Nyenhuis, J.A., Phillips, M., Shellock, F.G., Gonzalez-Martinez, J., & Rezai, A.R. (2004). Evaluation of specific absorption rate as a dosimeter of MRI-related implant heating. *Journal of Magnetic. Resonance. Imaging., 20*, 315–320.

Duru, F., Luechinger, R., Scheidegger, M.B., Luscher, T.F., Boesiger, P., & Candinas, R. (2001, January). Pacing in magnetic resonance imaging environment: Clinical and technical considerations on compatibility. *Eur Heart J., 22*(2), 113–124.

EEEI-C95.3. (1991). Recommended practice for measurements and computations with respect to human exposure to radiofrequency electromagnetic fields, 100 kHz to 300 GHz. *Institute of Electrical and Electronics Engineers (EEEI).*

Erlebacher, J.A., Cahill, P.T., Pannizzo, F., & Knowles, R.J. (1986). Effect of magnetic resonance imaging on DDD pacemakers. *Am. J. Cardiol., 57*, 437–440.

Hayes, D.L., Holmes, D.R. Jr., & Gray, J.E. (1987). Effect of 1.5 tesla nuclear magnetic resonance imaging scanner on implanted permanent pacemakers. *J. Am. Coll. Cardiol., 10*, 782–786.

Kanal, E., Borgstede, J.P., Barkovich, A.J., Bell, C., Bradley, W.G., Etheridge, S., et al. (2002). American College of Radiology white paper on MR-safety. *AJR Am. J. Roentgenol., 178*, 1335–1347.

Luechinger, R., Duru, F., Scheidegger, M.B., Boesiger, P., & Candinas, R. (2001). Force and torque effects of a 1.5-Tesla MRI scanner on cardiac pacemakers and ICDs. *Pacing Clin. Electrophysiol*, 24, 199–205.

Luechinger, R., Duru, F., Zeijlemaker, V.A., Scheidegger, M.B., Boesiger, P., & Candinas, R. (2002). Pacemaker reed switch behavior in 0.5, 1.5, and 3.0 Tesla magnetic resonance imaging units: Are reed switches always closed in strong magnetic fields? *Pacing Clin. Electrophysiol.*, 25, 1419–1423.

Luechinger, R., Zeijlemaker, V.A., Pedersen, E.M., Mortensen, P., Falk, E., Duru, F., et al. (2005, February). In vivo heating of pacemaker leads during magnetic resonance imaging. *Eur. Heart. J., 26*(4), 376–383.

Niehaus, M., & Tebbenjohanns, J. (2001). Electromagnetic interference in patients with implanted pacemakers or cardioverter-defibrillators. *Heart, 86*, 246–248.

Nitz, W.R., Oppelt, A., Renz, W., Manke, C., Lenhart, M., & Link, J. (2001). On the heating of linear conductive structure as guide wires and catheters in interbentional MRI. *Journal of Magnetic Resonance Imaging, 13*, 105–114.

Rezai, A.R., Finelli, D., Nyenhuis, J.A., Hrdlicka, G., Tkach, J., Sharan, A., et al. (2002, March). Neurostimulation systems for deep brain stimulation: In vitro evaluation of magnetic resonance imaging-related heating at 1.5 tesla. *J. Magn. Reson. Imaging., 15*(3), 241–50.

Roguin, A., Zviman, M.M., Meininger, G.R., Rodrigues, E.R., Dickfeld, T.M., Bluemke, D.A., et al. (2004). Modern pacemaker and implantable cardioverter/defibrillator systems can be magnetic resonance imaging safe. In vitro and in vivo assessment of safety and function at 1.5 T. *Circulation, 110*, 475–482.

Shellock, F.G. (1992). Thermal responses in human subjects exposed to magnetic resonance imaging. *New York Academy of Sciences, 649*, 260–272.

Shellock, F.G., & Crues, J.V. III. (2002). MR-safety and the American College of Radiology white paper". *AJR Am. J. Roentgenol., 178*, 1349–1352.

Shellock, F.G., Tkach, J.A., Ruggieri, P.M., & Masaryk, T.J. (2003). Cardiac pacemakers, ICDs, and loop recorder: Evaluation of translational attraction using conventional ("long-bore") and "short-bore" 1.5- and 3.0-Tesla MR systems. *J. Cardiovasc. Magn. Reson., 5*(2), 387–397.

Sommer, T., Vahlhaus, C., Lauck, G., von Smekal, A., Reinke, M., Hofer, U., et al. (2000). MR imaging and cardiac pacemakers: In-vitro evaluation and in-vivo studies in 51 patients at 0.5 T. *Radiology*, 215, 869–879.

KEY TERMS

Fluoroptic® Thermometry: A trademark patented by Luxtron's researchers in the 1970s; the fluoroptic® thermometry determines the temperature of the sensor by measuring the decay time of the emitted light. It is a persistent property of the sensor that its decay time varies precisely with temperature.

Implantable Cardioverter Defibrillator (ICD): A device that is implanted under the skin of patients that are at risk of sudden cardiac death, due to ventricular fibrillation. The rudiments of cardiac arrhythmia detection and treatment are incorporated into the implantable device. The device was designed primarily to deal with ventricular fibrillation. Its current use has, however, extended to include atrial and ventricular arrhythmias, as well as the ability to perform biventricular pacing in patients with congestive heart failure, and to pace, should there be any marked bradycardia.

Magnetic Resonance Imaging (MRI): An imaging technique based on the principles of nuclear magnetic resonance.

Pacemaker (PM): A medical device designed to regulate the beating of the heart. The purpose of an artificial pacemaker is to stimulate the heart when either the heart's native pacemaker is not fast enough, or if there are blocks in the heart's electrical conduction system preventing the propagation of electrical impulses from the native pacemaker to the lower chambers of the heart, the ventricles.

RF Field: Electromagnetic field generated during MRI procedures. RF is absorbed and re-transmitted by hydrogen nuclei. Re-emitted energy is used to obtain the image. RF field is measured as Vm^{-1}. The RF frequency is related to the static field; in most commercial systems (1.5 T), the frequency is 64 MHz.

Specific Absorption Rate (SAR): A measure of the rate at which radio frequency energy is absorbed by the body when exposed to radio frequency electromagnetic field.

Static Field: Strong static magnetic field used in MRI systems to align the magnetization vector of the hydrogen nuclei. It is measured as Tesla (T). In most commercial MRI system, the static field is 1.5 T.

Multi-Dimensional Modeling in the Health Industry

Markus Belkin
RMIT University, Australia

Brian Corbitt
RMIT University, Australia

Konrad Peszynski
RMIT University, Australia

INTRODUCTION

Today's information-rich and knowledge-based business society relies heavily on information technology (IT) and information systems (IS) design to enable the business to operate effectively and create a competitive advantage. Firms must align their IS design and performance with the core business competencies and business goals of the firm. There are multiple paths toward this end, and inefficiencies and conflicts may arise when the firm's IS strategies diverge from the business goals. There is no difference in the health industry, where conflicts exist between IS infrastructure and development, and business goals. The existence of inflexible mainframe IS unable to support modern technology such as the Internet, telemedicine, wireless technology, and real-time management software has compromised the business goals and business development in the health vertical to the extent that it has now fallen behind other comparable knowledge industries.

Where reference is made to more cohesiveness among IS capability, independence of the IS department, and the alignment of business goals, there is no mechanism or detail given on how this is achieved. Grover and Segars (2005) claim that while there have been studies that examine the "what" questions in strategic information system planning (SISP), particularly concerning the issue of IS business alignment, there has been little on the "how" questions.

A multidimensional cohesive model for IS planning and measurement of IS effectiveness has been developed as a means to more integrated planning and a simpler but more realistic means of assessing the effectiveness of the IS in business. The multidimensional cohesive model is applied to the selection and implementation of an information system in the health industry. The implications this has on the health industry include the opportunity to change to a more efficient business structure, a means to implement a modern technology- (Web-) based IS and an inherent capacity for change management.

BACKGROUND

Strategic information system planning (SISP) has evolved in method and style over the last decade on the basis that it is important because it emphasizes the need to bring information technology (IT) to align with and sometimes influence the strategic direction of the firm (Grover & Segars, 2005). In rich IT environments, this has a recognized relevance to competitiveness. However, although much has been studied with respect to business and IT alignment, little research has been undertaken into the mechanisms of SISP, including process planning.

Grover and Segars (2005) examined the evolution and maturing of SISP from the early 1970s and made several important observations. These were later supported by other researchers such as Earl (1993) and Sabherwal and King (1995). They found that many studies focused on planning content, with particular interest in methods and measurement of alignment between business and IS strategy (Burn & Szeto, 2000; King, 1998). They observed that these studies did little to illuminate the organizational aspects of planning.

Early studies by Pyburn (1983), in an attempt to identify institutionalized planning dimensions, actions, and behaviors, made field observations that noted the existence of both a *rational/structured* process and a *personal-informal* process. Earl (1993) made similar observations when he distinguished SISP approaches based on the degree of rationality and adaptability built into the planning process. Earl (1993), however, noted a hybrid organizational system of planning that seemed to be more effective than the highly structured and less adaptable rational approaches. This observation was ratified by the work of Sabherwal and King (1995).

More recent studies by Segars (1997) and Segars and Grover (1998) described and measured planning process dimensions and found that hybrid systems tended to be more successful and seemed to apply generally to a variety of industries. Through their research, Grover and Segars (2005) identified six important process dimensions of SISP: comprehensiveness, formalization, focus, flow, participation, and consistency. These dimensions are robust in describing the SISP design and extend beyond the methodological-based and less-generalizable descriptions of planning.

Wang and Tai (2003) add to the dimensions for success in SISP with their work on organizational contexts, commenting that most process-oriented research has recommended using integration and implementation mechanisms while not considering the possible contingent effect of contextual factors. They suggest that this may lead to the planning system being less adaptable to various organizational contexts and therefore be overly deterministic.

Wang and Tai (2003) acknowledge that although their work is generally supported by empirical data, a theory of IS planning is currently lacking. Their results did, however, support the contention that IS planning is a rational-adaptive process, supporting the claims of Earl (1993) and Grover and Segars (2005).

The link between strategic performance and planning has been found to be inconsistent by Grover and Segars (2005) and Premkumar and King (1992). Some indicators suggested for assessment of IS effectiveness have been IS usage, user information satisfaction (UIS), quality of decision-making, productivity from cost/benefit analysis, and system quality (Ein-Dor and Segev, 1978). The most commonly favored factors have been IS use and UIS. However, because of a lack of a theoretical framework for placing UIS within the greater context of overall IS effectiveness, its relevance as a performance measurement has been questioned (Grover & Segars, 2005).

Grover and Segars (2005) argue that successful SISP should achieve alignment between IS and business strategy; analyze and understand the business and associated technologies, foster cooperation and partnership between managers and user groups, anticipate relevant events/issues within the competitive environment, and adapt to unexpected organizational and environmental change. This multidimensional conceptualization approach is supported by Delone and McLean (1992).

However, further research is needed in order to define the construct space for effectiveness criteria. Delone and McLean (1992, 2003) have initiated research to this end with their IS Success model. Their model consists of six interdependent constructs, including system quality, information quality, use, user satisfaction, individual impact, and organizational impact (Delone & McLean, 1998). The measure of overall success should combine individual measures from these constructs to create a comprehensive scheme for performance.

Grover and Segars (2005) have developed a theoretically based construct space for IS effectiveness that complements the IS Success of Delone and McLean (1992). Their construct model provides a means of cross-validating the IS Success model and introduces a relative standard used for assessing performance.

To build a complete picture of IS effectiveness, evaluation must be conducted from both a macro (organizational) and micro (individual) view. Such evaluation is necessary because IS supports individual as well as organizational decision-making and can also provide competitive advantage.

From the organizational effectiveness literature, Brewer (1983) argues that there are three types of evaluation: process, response, and impact. Process evaluation involves the assumption that organizational members work to ensure efficient use of resources when resources are limited. This assessment is based on user dependence on IS, user perceptions of system ownership, and the extent to which IS is disseminated throughout organizational administration and operating procedure.

Response evaluation assesses the individual or the organization to the IS service or product. This assessment has significance in respect to user resistance to innovation and implementation. Any resistance or habitualization must be identified to ensure successful implementation. This assessment also considers complex variables such as user's beliefs and attitudes toward IS in general, which are important for fulfilment of IS planning (Grover & Segars, 2005).

Impact evaluation represents the most comprehensive and most difficult to assess evaluation. It is associated with the direct effects of IS implementation on the individual and/or the organization.

Grover and Segars (2005) model produces six classes of IS effectiveness measurement that define the overall construct space for IS effectiveness. As shown in Figure 1, the evaluation of IS is initiated by choice of the relevant evaluative referent. The first three classes of effectiveness measures are associated with macro (organizational) evaluation.

From their empirical work developing this model, Grover and Segars (2005) state, "It seems that both theoretical depictions strongly imply that IS effectiveness is multidimentional in terms of types of measures and level of analysis" (782). This supports their earlier stated contention and supports the argument by other authors (Earl 1993; Pyburn 1983; Sullivan 1985).

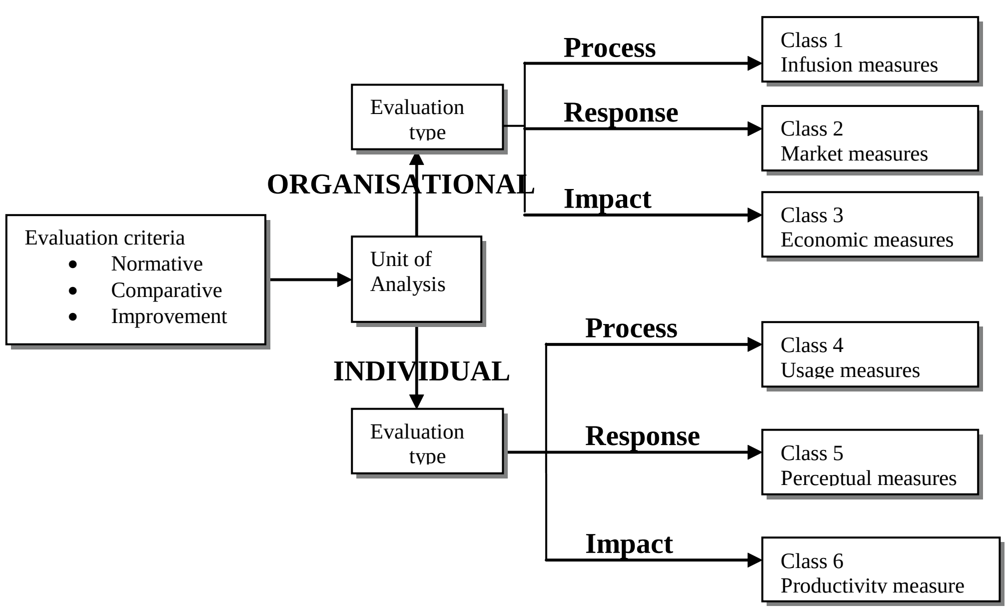
Further research by Grover and Segars (2005) on evaluation perspective raises questions asking from whose perspective the evaluation is being carried out. Although IS may be viewed as effective from one standpoint, it may be viewed as the opposite from another. Cameron and Whetten (1983) suggest that one reason no "best" criterion exists is because there is no "best" constituency. As such, Grover and Grover (2005) state that for the evaluation of IS effectiveness, the specific views of all groups should be considered because they help to increase awareness of the value of the IS and help the understanding of the multidimensionality of IS effectiveness.

MAIN FOCUS

The literature on the approach to measurement of IS effectiveness shows a great variation in the measurement techniques and the possibility of inefficiencies in the deployment of effectiveness measures through personal agendas or bias in the management team. Several authors exploring SISP, such as Pyburn (1983), Earl (1993), and Sabhewal and King (1995), agree that the planning process is most successful when rational and adaptive pathways are used in the design process. However, there is no mechanism suggested for possible pathways for this to happen. It is believed that there are several other shortcomings in the planning models presented in business: IT misalignment, no consideration for team member selection in either top-down or bottom-up situations, no consideration as to how people communicate to make plans (i.e., knowledge management/extraction and organizational learning), clear definition of business goals by thorough business analysis involving stakeholders from management to end-users required, and ensuring that the current IS hardware/software has the capability to handle the planned IS changes.

The literature cited on effectiveness measure is in a more confused state with so many criteria for measurement quoted. These criteria are clearly unidimensional and seem to be dissociated from what SISP is attempting

Figure 1. The construct space for IS effectiveness (Grover & Segars, 2005)



to achieve; that is, the original stated business goal(s) conjoined with the IS pathway during the SISP. The models presented by Grover and Segars (2005) and Wang and Tai (2003) are in themselves a demonstration of inadequacy. The arrows in Figure 1 representing direction and pathways are represented as being unidirectional. This implies that the models have a start and a finish. They are static, single event models. They would therefore be inconsistent with ongoing support and fine-tuning of SISP and maintenance of any competitive advantage gained with first implementation. Therefore, an alternative model is proposed based on a case study in medical pathology in Australia.

The case study is presented at two levels: the overall pathology practice and a specific pathology department. In the overall pathology practice, there was a merger of two pathology practices, one large and one medium into one practice. The workload for staff doubled overnight. The IT system of the larger practice was retained as the new group's IT platform. This was in theory totally inadequate and was proven to be so in practice in a short period of time post-merger. No business analysis premerger was performed to assess the existing system's capability to cope with double the workload. The IT platform consisted of an older mainframe computer that ran in-house developed software. The software was written in a totally unsuitable language for scientific applications (COBOL), and the storage capacity of the system was such that records could only be held on the system for one month before having to be transferred to Microfiche. The system was, in fact, solely a database and had no laboratory functionality to capitalize on the technology capability of the practice's analysers (applications that assist with quality control data and graphs, diagnostic graphics, work lists, and management statistics, and utilize LANs). This greatly compromised the functionality of the laboratory and the service to referring medical practitioners. There were personal agendas (software designer's claim: "We are the bigger laboratory so we know best") and cost considerations that determined the decision to stay with this unsuitable antiquated IT platform.

At the practice department level, the hematology department used analyses that were capable of extensive data and graphics generation for both diagnostic and quality control purposes. Each analyzer had a PC that acted as a controller for the analyzer and a data/graphics generator. The ability to disseminate the graphics would render the department paperless, saving time and consumables and increasing efficiency considerably. Incumbent mainframe systems were unable to accommodate this facility, and unless a supplementary graphics capable IT system was implemented, this great opportunity for efficiency improvement and consumable cost reduction would be lost. Management data from mainframe systems is statistically based on providing information as requested on types and numbers of tests performed, the number of tests from each referring doctor, and workstation performance. Enhancement of computer-based management performance, such as roster generation, reagent tracking, and supply chain management in real time, is not possible on existing systems due to lack of graphics capability. Many management functions are therefore still performed manually by department senior staff.

As in the example of the overall laboratory case, the drivers for decisions for choice of analyzers and IT systems are still made by top-level management with little consultation with end-users, and are cost based. The situation of the IT and business misalignment cited previously (Grover & Segars, 2005) is applicable, and the agendas also embrace self-preservation of employment.

Both examples presented in the case study show how inadequate planning (SISP) and personal agendas have and are still compromising the efficiency and adaptability to modern technological change in medical pathology. This is believed to affect the bottom line. Continuation with existing IT infrastructure in medical pathology is broadening the gap with modern technology and increasing the degree of difficulty for change. The realization of stakeholders of proper SISP and criteria for assessment of the effectiveness of IS other than cost considerations applies to medical pathology as rigorously as it does to any business. A change in attitude toward SISP and the adoption of processes based on the model proposed in this chapter could revolutionize the medical pathology industry by facilitating a major change in its business infrastructure from a pyramidal hierarchy to laterally linked self-funded business units. The models proposed in this chapter offer a mechanism for enhanced SISP and measurement of IS effectiveness, leading to the development of a cohesive business model.

The Cohesive Business Model (CBM) comprises two overlapping and coexisting units. The change source consists of two main functions. It is the innovation part of the CBM as it relies on the interaction of experienced people to capture ideas, knowledge through experience, and historical data. Through careful business analysis, the ideas and innovations are formatted into plans, models, and processes to carry them to implementation. The plans and models define team selection and project management as well as an absolute definition of the business goal(s) in the SISP. This process then provides the firm with an inherent facility for change management, as represented in Figure 1.

The process source also consists of functions. It is the integration part of the CBM as the models and processes from the change source are integrated into the business and the business change is effected. The end users provide feedback to the innovators, which give information regarding client issues as the perceptions are at the coal face. The process source unit is where quality management takes place through interaction and feedback at the client interface.

The change source (innovation side) of the CBM works closely with the IS department. This relationship differs from what is a common situation cited in the literature where the IS department is dominant in the relationship with the business managers. This model is driven by business and recognizes the IS department as a valuable, essential partner to advise on the best way to integrate the latest IT developments into the CBM. The IS department's facility for development of software and technical backup is defined in the plans, models, and processes of the innovation. The involvement of the IS department, together with the innovation and project management teams in staff training and alpha and beta testing of the developed IS, is imperative to the success of that project.

The CBM is multidimensional and is supported by provision for assessment of internal and external factors that may affect the implemented model in a positive or negative way. The links between components of the CBM, unlike other models cited in the literature, are bidirectional. The CBM is dynamic and fluid and has the facility to respond to potential influences quickly and effectively to ensure that the business maintains its functionality and competitive advantage.

FUTURE TRENDS

A review of literature on Laboratory Information Systems (LIS) and laboratory management systems relative to the case study in this article acknowledges a need for improvement in effectiveness and efficiency in clinical laboratories. In efforts to achieve this, several approaches by authors have been taken. These include the functional process approach of Goldschmidt, deVries, van Merode, and Derks (1998), the accountant's perspective of Revere (2004), the cost analysis approach of Mayer (1998), and the LIS requirements for the future work of Bender and McNair (1996). Bender and McNair's (1996) work supports and promotes the use of open architecture systems, which implies that the market will develop modular, scalable, and cost-effective LIS without the dependence on individual manufacturers and hardware/software systems that characterize current systems.

There is a marked diversity in these approaches with perhaps only the work of Bender and McNair (1996) considering the LIS as a whole. Their approach is truly significant when considering the IS management misalignment problems cited earlier in this article. Our planned future research centers on the implementation of a modular real-time management system designed by the researchers for the clinical laboratory using principles of the CBM. The possibility that the system will enable laboratories to change their business infrastructure as well as improve efficiency, introduce new technologies, and reduce costs will be investigated. The expected positive effect this approach will have on IS/business misalignment will also be investigated.

CONCLUSION

SISP is recognized as an event that results in the building of an IS to support a business goal. The ability of the IS to attain that business goal must surely be the measure of the effectiveness of the SISP. Many of the quoted measures of effectiveness of SISP and IS should be regarded as part of the planning brief so they are inherently achieved by the IS during the processes of alpha- and beta-testing, the user friendliness, usage, cost/benefits, and ROI.

The adoption of the alternative model and its approach to SISP introduces a multidimensional approach to SISP. The multidimensional approach is bidirectional in that it allows for an inherent change management facility in the SISP, which keeps the process abreast of changing internal and external influencing factors to the benefit of the firm. Competitive advantage is maintained through more rapid response to these changes. Future research will test the model presented in this article.

The multidimensional approach to SISP also allows for more standardization in design process by defining what is a design component of IS and what is an indicator of effectiveness of the IS. More research needs to be undertaken to fully elucidate these definitions to further enhance our understanding of this complex environment.

REFERENCES

Bender, J., & McNair, P. (1996). User requirements on the future laboratory information systems. *Computer Methods and Programs in Biomedicine*, *50*, 87–93.

Brewer, G.D., (1983). Assessing outcomes and effects. In K.S. Cameron, & D.A. Whetten (Eds.), *Organisational effectiveness: A comparison of multiple models*. San Diego, CA: Academic Press.

Burn, J.M., & Szeto, C. (2000). A comparison of the views of business and IT management on success factors for strategic alignment. *Information & Management, 37*(4), p197–216.

Cameron, K.S., & Whetten, D.A. (1983). Organisational effectiveness: One model or several? In K.S. Cameron, & D.A. Whetten (Eds.), *Organisational effectiveness: A comparison of multiple models*. San Diego, CA: Academic Press.

Delone, W.D., & McLean, E.R. (1992). Information systems success: The quest for the dependent variable. *Information Systems Research*, *3*(1), 60–95.

Delone, W.D., & McLean, E.R. (2003). The Delone and McLean model of information systems success: A ten-year update. *Journal of Management Information Systems*, *19*(4), 9–30.

Earl, M.J. (1993). Experiences in strategic information systems planning. *MIS Quarterly, 17*(1), 1–24.

Ein-Dor, P., & Segev, E. (1978). Organisational context and the success of MIS. *MIS Quarterly, 24*(10), 1–24.

Gerwin, D., & Kolodny, H. (1992). *Management of advanced manufacturing technology: Strategy, organisations and innovation*. New York: Wiley.

Goldschmidt, H.M.J., de Vries, J.C.M., van Merode, G.C., & Derks, J.J.M. (1998). A workflow management tool for laboratory medicine. *Laboratory Automation and Information Management, 33*, 183–197.

Grover, V., & Segars, A.H. (2005). An empirical evaluation of stages of strategic information systems planning: Patterns of process design and effectiveness. *Information & Management, 42*, 761–799.

King, W.R. (1988). How effective is your information system planning? *Long Range Planning*, *21*(5), 103–112.

Mayer, M. (1998). Laboratory cost control and financial management software. *Clinica Chimica Acta, 270*, 55–64.

Prekumar, G., & King, W. R. (1992). An empirical assessment of information systems planning and the role of information systems in organisations. *Omega 25*(1), 107-121.

Pyburn, P.J. (1983). Linking the MIS plan with corporate strategy: An exploratory study. *MIS Quarterly, 7*(2), 1–14.

Revere, L. (2004). Improving resource efficiency through management science. *Journal of Healthcare Management, 49*(5).

Sabherwal, R., & King, W.R. (1995). An empirical taxonomy of the decision making process concerning strategic applications of information systems. *Journal of Management Information Systems, 11*(4), 177–214.

Segars, A. H. (1997). Assessing the unidimensionality of measurement: A paradigm and illustration within the context of information systems research. *Omega*, *25*(1), 107-121.

Segars, A.H., & Grover, V. (1998). Strategic information systems planning success: An investigation of the construct and its measurement. *MIS Quarterly, 22*(2), 139–163.

Sullivan, C. H. (1985). Systems planning in the information age. *Sloan Management Review, 26*(2), 3-12.

Wang, E.T.C., & Tai, J.C.F. (2003). Factors affecting information systems planning effectiveness: Organisational contexts and planning systems dimensions. *Information & Management, 40*, 287–303.

KEY TERMS

Bidirectional Modeling. A business or functional model in which data are exchanged to and from all components of the model in real time.

Cohesive Business Model: A multidimensional business model developed for information systems planning and measurement of IS effectiveness based on the clearly defined business goals of the firm and with an inherent capacity for change management. This model is fluid, dynamic, and ongoing.

IS Effectiveness: The ability of an IS to meet the success criteria as determined by the firm.

IT Misalignment: A difference in the developmental pathways of the IT department and the business goals and competencies of the (same) firm.

Laboratory Information System (LIS): An IS in a medical laboratory for data and record collection and storage, results handling and dissemination, and functional statistics generation.

Multidimensional Modeling: A process of business modeling based on rational/structured and personal/informal processes that are fluid, dynamic, and ongoing. This process considers the relationship among all components of the model bidirectionally and continuously.

Strategic Information System Planning (SISP): The planning of an information system based on many described and measured processes and aligned with the business goals and competencies of the firm undertaken to increase the competitive advantage of the firm.

Unidimensional Modeling: A process of business modeling based on rational/structured and personal/informal processes that are finite; it has a beginning and an end. This process considers the relationship among all components in one direction only and is a static process.

M

Myoelectric Control of Prosthetic Devices for Rehabilitation

Rami N. Khushaba
University of Technology, Australia

Adel A. Al-Jumaily
University of Technology, Australia

INTRODUCTION

Bio-signals patterns analysis problems have enjoyed a rapid increase in popularity in the past few years. The electromyography (EMG) signal, also referred to as the Myoelectric signal (MES), recorded at the surface of the skin, is one of the biosignals generated by the human body, representing a collection of electrical signals from the muscle fibre, acting as a physical variable of interest since it first appeared in the 1940s (Scott, 1984). It was considered to be the main focus of scientists, and was advanced as a natural approach for the control of prosthesis, since it is utilising the electrical action potential of the residual limb's muscles remaining in the amputee's stump (which still has normal innervations, and thus is subject to voluntary control) as a control signal to the prosthesis—in other words, it allows amputees to use the same mental process to control their prosthesis as they had used in controlling their physiological parts; however, the technology in that time was not adequate to make clinical application viable. With the development of semiconductor devices technology, and the associated decrease in device size and power requirements, the clinical applications saw promise, and research and development increased dramatically.

This control approach referred to as myoelectric control has found widespread use for individuals with amputations or congenitally deficient limbs. The methodologies used in this approach of control spanned the range from classical multivariate statistical and syntactic methods, to the newer artificial intelligence (symbolic and connectionist) approaches to pattern processing. In these systems, voluntarily controlled parameters of myoelectric signals from a muscle or muscle group are used to select and modulate a function of a multifunction prosthesis. The essential elements of a myoelectric control of prosthesis devices are shown in the block diagram schematic of Figure 1 (Merletti & Parker, 2004). The Myoelectric control system is based on the noninvasive interfaces designed for casual wear.

In the following sections, the details and related works to the research problem are given; after that, the methodology adopted is explained and discussed, followed by future work, and finalised with conclusion.

BACKGROUND AND RELATED WORK

Continuous myoelectric-controlled devices are one of the challenging research issues, in which the prosthetic is controlled in a manner proportional to the level of myoelectric activity. Although the success of fitting these systems for single device control is apparent, the extension to control more than one device has been difficult (Hudgins, Parker, & Scott, 1993), but, unfortunately, it is required for those with high-level (above the elbow) limb deficiencies, and the individuals who could stand to benefit from a functional replacement of their limbs (Englehart, Hudgin, & Parker, 2001). It has been proved that the MES signal exhibits a deterministic structure during the initial phase of muscle contraction, as shown in Figure 2 for four types of muscle contractions measured using one surface electrode. Based on this principle, a continuous myoelectric control strategy based on the use of pattern classifiers was the main focus during the last years for most scientists in related fields.

The MES patterns exhibit distinct differences in their temporal waveforms. Within a set of patterns derived from the same contraction, the structure that characterizes the patterns is sufficiently consistent to maintain a visual distinction between different types of contraction. Hudgins et al. (1993), and Lighty, Chappelly, Hudgins, and Englehart (2002) aligned the patterns using a cross-correlation technique, and showed that the ensemble average of patterns within a class preserves this structure. The myoelectrical signal is essentially a one-dimensional pattern, and the methods and algorithms developed for pattern recognition can be applied to its analysis. The myoelectric control, system-based pattern classifier consists of four broad system components, which are (Ciaccio, Dunn, & Akay, 1993; Lusted & Knapp, 1996) :

- Myoelectric signal acquisition using surface or implanted electrodes, mostly surface electrodes.
- Signal conditioning and features extraction.
- Pattern recognition algorithms to classify the signal into one of multiple classes.
- Mapping of classified patterns to interface actions that control external devices.

A general look onto the myoelectric control system components reveal that it operates at a few stages of machine pattern recognition or interpretation for biosignals that were proposed in Ciaccio et al. (1993), and Lusted and Knapp (1996). Also, specific features are extracted usually because the motivation is toward the evaluation of myoelectric signal features in ways which are not tied to accurate estimates of signal characteristics, but rather to the intrinsic quality of the features as control signals for a desired device, which have been usually a prosthetic hand robot for the purpose of rehabilitation.

The features extraction is considered as the most important part, because the success of any pattern classification system depends almost entirely on the choice of features used to represent the continuous time waveforms (Hudgins et al., 1993). Although the

Figure 1. Normal and myoelectric control system

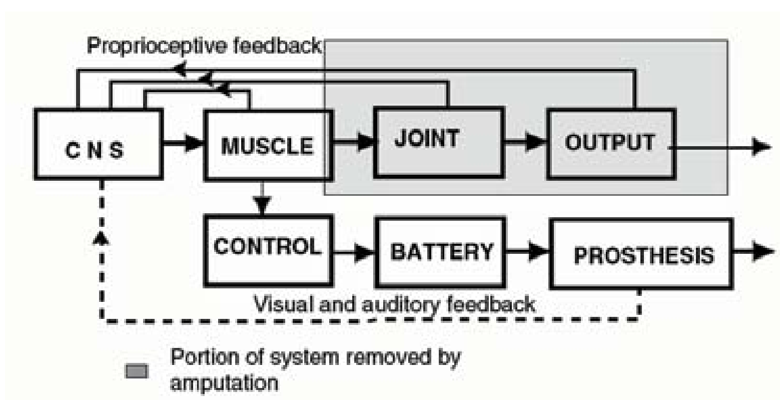
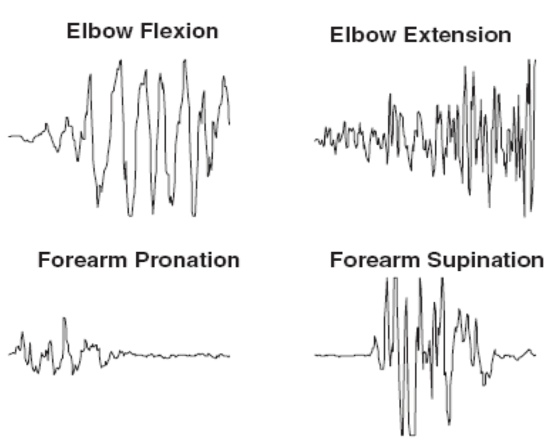


Figure 2. Patterns of transient MES activity recorded using a single bipolar electrode pair, placed over the biceps and triceps



literature includes many papers which explore the extraction of features from the MES for controlling prosthetic limbs, there have been few works which make quantitative comparison of their quality. Overall, a high-quality MES feature space should have the following properties (Boostani & Moradi, 2003):

- **Maximum class separability:** A high-quality feature space is that which results in clusters that have maximum class separability or minimum overlap. This ensures that the resulting misclassification rate will be as low as possible.
- **Robustness:** The selected feature space should preserve the cluster separability in a noisy environment as much as possible.
- **Complexity:** The computational complexity of the features should be kept low, so that the related procedure can be implemented with reasonable hardware, and in a real-time manner

Raw MES offers us valuable information in a particularly useless form. This information is useful only if it can be quantified. Various signal-processing methods are applied on raw EMG to achieve the accurate and actual MES. Most of the researches on myoelectric control focused on the extraction of features that will best serve the purpose of controlling the prosthetic arm; the first attempts were made since the 1980s, when Doerschuk (Doerschuk, Gustafson, & Willsky, 1983) and Graupe (Graupe, Salahi, & Zhang, 1985) (that were both based on Graupe and Cline (1975)), used the parameters of some stochastic models, such as an autoregressive (AR) model, or auto regressive moving average (ARMA) model as features set for the myoelectric control system. Auto-regressive parameters (AR) were used as features for the MES signal (Karlik, Pastaci, & Korurek, 1994; Kocyigit, Karlik, & Korurek, 1996), as AR parameters can accurately estimate the power spectrum of the signal. It is worthwhile to observe some important advantages of modelling the signal in this way (Lamounier, Soares, Andrade, & Carrijo, 2002):

- Variations in the positioning of the electrodes on the surface of the muscle do not severely affect the AR-coefficients.
- The amount of information to be presented to the classifier is greatly reduced. Therefore, the total processing time is also reduced.

Recently, all attempts to extract features from the MES can be classified, in general, into two categories (Vuskovic, Pozos, & Pozos, 1995), although there are still many using the AR model for features extraction, and those two categories are:

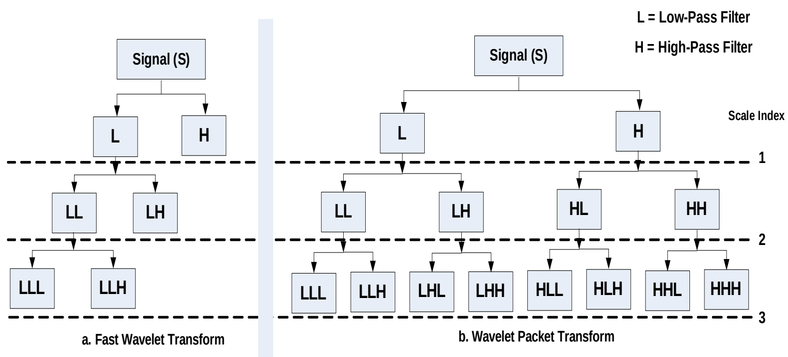
- **Temporal approach:** Identifies the attributes of the raw MES signal that characterize its temporal structure relative to a specific muscular function.
- **Spectral approach:** Uses information contained in frequency domain, which leads to a better solution for encoding the MES signal.

It was shown through literature that the spectral approach results were superior to that of the temporal approach, and that is why we adopted a spectral technique for features extraction.

METHODOLOGY

Due to the nonstationary behaviour of biosignals, the use of time frequency representations is highly desirable with the view to deriving meaningful features. These representations provide the means to regard the phenomenon simultaneously under two different points of view. The frequency characteristics, as well as the temporal behaviour, can be described with respect to the uncertainty principle. With multiresolution analysis, fast wavelet transform (FWT) leads to dyadic pyramidal implementation using filter banks and the corresponding Mallat algorithm (Mallat, 1989). FWT develops the two channel filter banks, through which the signal is split into two subspaces, L and H, which are orthonormally complementary to each other, with L being the space that includes the low-frequency information about the original signal, and H includes the high-frequency information. We keep repeating the decomposition of the low frequency subspace L. Compared to FWT, the wavelet packet (WP) method is a generalization of wavelet decomposition that offers a richer range of possibilities for signal analysis. The wavelet packet transform (WPT) not only decomposes the approximation coefficients, but also the details coefficients H as shown in Figure 3, in which the level of decomposition is to scale Index 3. The WPT provides a selective subbanding, which is the best for

Figure 3. Tree-like structure of digital signal processing by



Note: a) FWT analysis to scale 3; b) WPT analysis to scale index 3

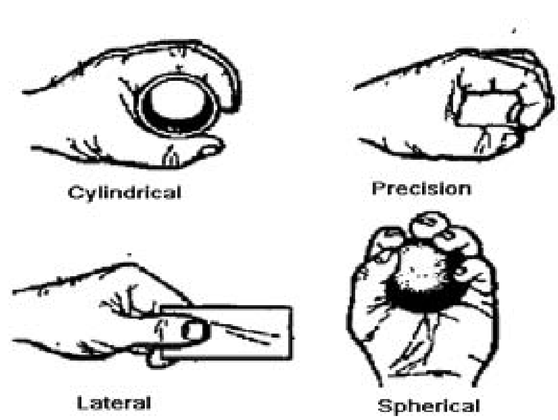
a given part of signals selected, such as in the case of the MES in which the most signal energy is positioned in between 35Hz–250Hz (Chan, Yang, Lam, Zhang, & Parker, 2000; Hudgins et al., 1993). The selective banding in WPT is determined by the order of low-pass and high-pass filtering from a tree-like structure using these filters (5).

Our proposed approach uses wavelet packet transform (WPT) for features extraction, with four levels of decomposition of the same database that is used by (Sijiang & Vuskovic, 2004) from the San Diego State University (focusing on the synergetic control of groups of finger joints that correspond to the basic prehensile motion), in which the MES signals of the four channels are sampled and recorded as the raw signal data. There are six categories of grasp types: small cylinder (SC), large cylinder (LC), small ball (SB), large ball (LB), small disk (SD), and a key (SK) as shown in Figure 4. Each of the six categories has 30 recorded grasp instances, respectively. There are 180 total recorded grasp instances, and the recorded data are used for feature extraction. Each instance generates a feature vector as the result of feature extraction by the WPT method. Four types of grasps are considered in this database, those are: cylindrical grasp, precision grasp, lateral key grasp, and spherical grasp, as shown in Figure 4. The input feature vectors are computed in two simple steps—first, each record of the database is decomposed using WPT with the fourth level (using two wavelets families: Daubechies, and Symmlet); second, features are extracted by simply determining every component of the feature vector ($f_1, f_2, ..., f_N$) using the Euclidean norm. Assuming the number of nodes in the fourth level is simply N , and that each node contains M elements, representing a row vector r_j for $j = 1, 2, ..., N$ the Euclidean norm is given by:

$$\left[f_j := \left\| r_j \right\|_2 = \sqrt{\sum_{i=1}^{M} v_i^2} \right] \qquad (1)$$

This means that each feature f_i is determined as the square root of the energy of the wavelet coefficients in the corresponding node, representing time and frequency information of the specific signal. A single feature especially describes a certain frequency range, which is equal to that described by wavelet coefficients underlying this feature. The method of computing the energy of wavelet coefficients was introduced in

Figure 4. Four grasp types



Pittner and Kamarthi (1999), but they used the FWT instead of WPT.

Because WPT possesses the useful property of identifying distinguishing characteristics from the original signals, we map the original signal into many WPT feature space. The number of valid decompositions increases exponentially with the number of decomposition level, so a problem found here is to compute the optimal decomposition. In our experiment, we found that the fourth level was very appropriate to deal with such a problem, as further decomposition is not necessary for our work. The number of features selected at the fourth decomposition level was also a matter of study for us, were we started with 12 features , and gradually increased the number of features extracted per channel (till 16 feature/channel), reaching to optimal results. The other part of our system was the classifier—to distinguish the six classes we have of movements in the database, we simply choose the Linear Discriminant Analysis (LDA), and achieved a high accuracy from the proposed system. The block diagram of the proposed system for classifying six categories of movements is shown in Figure 5, simply containing the extraction of features from the energy of WPT and followed by the LDA classifier.

The proposed system is different from that proposed in Englehart et al. (2001) by Englehart and his colleagues, where they used records of data each with 256 sample, and later extracted features by WPT, due to the size of their feature vectors they used Principle Component Analysis to reduce the feature set dimension, and followed that was an LDA part for classification, achieving an average error rate of 0.5% for four classes of motion, and 2% for six classes problem. In our case, we used the records from SDSU, each containing 400 samples, and extracted a number of features from the Euclidean norm of wavelet coefficients, achieving 64 features for the whole four channels, thus not requiring PCA, and conveyed those to an LDA classifier, and the results were very successful, as shown in Table 1-1, consisting of the number of features per channel, the error rate, and the wavelet families that we used in our experiments.

Table 1. Practical results from our experiment

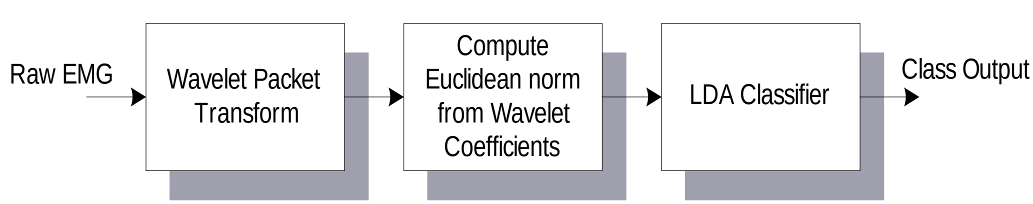
Features/Channel	Error %	Wavelet
16	0.56%	Symmlet
14	0.56%	Symmlet
12	1.11%	Symmlet
16	0.52%	db4
14	1.67%	db4
12	2.78%	db4

It is obvious from the table of experimental results that the case of 16 features per channel for Symmlet family of WPT, and Daubechies of WPT gave the best results; also, increasing the decomposition level to five gave 100% accuracy for the both families of wavelet. Our design was used to classify six classes of motion, achieving an error rate of 0.56% using Symmlet family, whereas in Englehart et al. (2001), the six classes problem reached only 2% of error rate.

FUTURE TRENDS

The fields of human-computer interaction and robotics emphasise the necessity of humanizing machine interaction, thus calling for more intuitive interfaces. The biocontrol systems technology had long served to provide the support required for people with paralysis, resulting from many causes, by focusing on noninvasive interfaces (those designed for casual wear). The EMG signal was utilised mainly in two applications in the last few years: those wearing controlling prosthetics devices for rehabilitations of patients, and in speech recognition that emerged recently. In both of the applications, the EMG signals were used to build a successful natural

Figure 5. Block diagram of the proposed system



controller with accurate results. Work in this field is a continuing one to achieve optimum results. Various methods have been utilised in literature in this field, but no one yet have used Type 2 fuzzy sets in myoelectric control; that is expected to achieve great results, as fuzzy set Type 2 is the extension to the fuzzy logic that has been used in this field, providing more accurate results than those based to statistical methods.

CONCLUSION

In this article, a simple and very effective wavelet-based approach has been explained for the problem of MES classification for the purpose of rehabilitation of patients after stroke. A simple method of extracting features from the wavelet coefficients was adopted and proven to be very effective in increasing the classification accuracy of the designed system. Although being simple, the accuracy of the classification system reached 99.44% for the case of six classes of motion, using the Symmlet family of the order five. The design described in this article represented a new promising approach for practical implementation of Myoelectric-controlled prosthetics, which we will be further enhancing for more applications, where we are currently investigating fuzzy approaches in feature selection as future work, and the application of artificial intelligence in pattern recognition rather than the LDA classifier.

ACKNOWLEDGMENT

The authors would like to acknowledge the support of Dr. Marko Vuskovic from the San Diego State University for supplying the database.

REFERENCES

Boostani, R., & Moradi, M. H. (2003). Evaluation of the forearm EMG signal features for the control of a prosthetic hand. *Physiological Measurement, 24*, 309–319.

Chan, F. H. Y., Yong-Sheng Y., Lam, F. K., Zhang, Y. T., & Parker, P. A. (2000). Fuzzy EMG classification for prosthesis control. *IEEE Transactions on Rehabilitation Engineering, 8*(3), 305–311.

Ciaccio, E. J., Dunn, S. M., & Akay, M. (1993). Biosignal pattern recognition and interpretation systems. *Engineering in Medicine and Biology Magazine, IEEE, 12*(3), 89–95.

Doerschuk, P. C., Gustafson, D. E., & Willsky, A. S. (1983). Upper extremity limb function discrimination using EMG signal analysis. *IEEE Transactions on Biomedical Engineering, BME-30*(1), 18–29.

Englehart, K., Hudgin, B., & Parker, P. A. (2001). A Wavelet-based continuous classification scheme for multifunction myoelectric control. *IEEE Transactions on Biomedical Engineering, 48*, 302–311.

Graupe, D., Salahi, J., & Zhang, D. (1985). Stochastic analysis of myoelectric temporal signatures for multifunction single-site activation of prostheses and orthoses. *Journal of Biomedical Engineering, 7*(1), 18–29.

Graupe, J., & Cline, K. (1975). Functional separation of EMG signal via ARMA identification methods for prosthesis control purposes. *IEEE Transactions on Systems, Man, and Cybernetics, SMC, 5*, 252–259.

Hudgins, B., Parker, P., & Scott, R. N. (1993). A new strategy for multifunction myoelectric control. *Biomedical Engineering, IEEE Transactions on, 40*(1), 82–94.

Karlik, B., Pastaci, H., & Korurek, M. (1994). Myoelectric neural networks signal analysis. In *Proceedings of the 7th Mediterranean Electrotechnical Conference* (Vol. 261, pp. 262–264).

Kocyigit, Y., Karlik, B., & Korurek, M. (1996). *EMG pattern discrimination for patient-response control of FES in paraplegics for walker supported using artifical neural network (ANN).* Paper presented at the 8th Mediterranean Electrotechnical Conference, MELECON '96.

Lamounier, E., Soares, A., Andrade, A., & Carrijo, R. (2002). A virtual prosthesis control based on neural networks for EMG pattern classification. *6th IASTED International Conference on Artifical Intelligence and Soft Computing (ASC)* (pp. 425–430). Banff, Alberta, Canada.

Lighty, C. M., Chappelly, P. H., Hudgins, B., & Englehart, K. (2002). Intelligent multifunction myoelectric control of hand prostheses. *Journal of Medical Engineering & Technology, 26*(4), 139–146.

Lusted, H. S., & Knapp, R. B. (1996). Controlling computers with neural signals. *Scientific American*, 82–87.

Vuskovic, M. I., Pozos, A. L., & Pozos, R. (1995, October 22–25). *Classification of grasp modes based on Electromyographic patterns of preshaping motions.* Paper presented at the Proceedings of the Internet, Conference on Systems, Man and Cybernetics, Vancouver, B.C., Canada.

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 11*(7), 674–693.

Pittner, S., & Kamarthi, S. V. (1999). Feature extraction from wavelet coefficients for pattern recognition tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 21*(1), 83–88.

Merletti, R. & Parker, P. (2004). Electromyography physiology, engineering, and noninvasive applications. *IEEE Press Engineering in Medicine and Biology Society.*

Scott, R. N. (Ed.). (1984). *An introduction to myoelectric prosthesis*. The Bioengineering Institute, University of New Brunswick.

Sijiang, D., & Vuskovic, M. (2004). Temporal vs. spectral approach to feature extraction from prehensile EMG signals. In *Proceedings of the 2004 IEEE International Conference on Information Reuse and Integration* (pp. 344–350).

KEY TERMS

Biocontrol System: A mechanical system that is controlled by biological signals—for example, a prosthesis controlled by muscle activity.

Electromyogram (EMG): These are electrical potentials arising from muscle movements. They originate from the motor cortex when the brain sends action potentials along appropriate nerve tracts. They then transmit to the muscle groups, resulting in contractions of the muscle fibers, and are accessible as bioelectric signals under direct volitional control.

Grasp: To take hold of or seize firmly with, or as if with the hand.

Myoelectric Control: A man-machine control scheme in which myoelectric signals are used as control signals.

Noninvasive Interfaces: Those interfaces that acquire signals transcutaneously, using surface electrodes, which are preprocessed to reduce noise content.

Rehabilitation: The process of restoration of skills by a person who has had an illness or injury, so as to regain maximum self-sufficiency, and function in a normal—or as near normal—manner as possible. For example, rehabilitation after a stroke may help the patient walk again and speak clearly again.

Wavelet Transform: In mathematics, wavelets, wavelet analysis, and the wavelet transform refers to the representation of a signal in terms of a finite length or fast-decaying, oscillating waveform (known as the mother wavelet). This waveform is scaled and translated to match the input signal. In formal terms, this representation is a wavelet series, which is the coordinate representation of a square integrable function with respect to a complete, orthonormal set of basis functions for the Hilbert space of square integrable functions.

Nanorobotics: Applications in Bionanotechnology

Sourabh Bansal
Illinois Institute of Technology, USA

INTRODUCTION

Nanorobotics is a young and a challenging discipline of nanotechnology that is related to the design and simulation of robots or nonbiological machines on the scale of a nanometer (10^{-9} meter/1billionth of a meter). *Nanorobotics* is also termed as the science of studying nanomachines, commonly known as Nanorobots (nanobots or nanoids), which are in the range of 0.1-10 micrometres[1].

BACKGROUND

"There's Plenty of Room at the Bottom"[2]

This was the title of the first talk on nanotechnology given by the famous physicist, Nobel Prize winner Richard Feynman, in 1959.

Feynman offered $1000 to anyone who could either build a motor that would fit inside a 1/64" x 1/64" x 1/64" box, or write a page of text with letters that are small enough for the Encyclopedia Britannica to be printed on a pin head. As soon as Feynman announced these two prizes, the first one got claimed right away. Bill McLellan, who had really good hands, made a small motor almost immediately to fit inside the box of 1/64" x 1/64" x 1/64" without creating any new technology. However, the second Feynman prize did not get claimed until very late (1985). Tom Newman and Fabian Pease (Stanford University) used electronic beam lithography technique to write *A Tale Of Two Cities* at the length scale requested by Feynman. Hence, based on Richard Feynman's vision of miniature factories using nanomachines to build complex products, "nanotechnology" evolved over the years as the nascent, yet one of the most promising, fields of technology.

Nanotechnology is an extremely diversified field with its applications in numerous areas including medicine, chemistry and environment, information and communication, energy, and consumer goods. Out of these areas, medicine or life sciences is perhaps the most targeted application of nanotechnology. This area of study of nanotechnology in the field of biology and medical sciences is coined as *Bionanotechnology* or *Nanobiotechnology*. Nanotechnology led to the development of the novice field of Nanorobotics, which came into existence in 1980s. This area of study was majorly focused towards the generation of machines at nanoscale that would be capable of manipulating the nanoparticles.

In 1986, Dr. K. Eric Drexler[3] published his research on nanosystems, in which he discussed a field that would be derived largely from the macroscopic robots. Later, this field segregated into two parallel categories: one comprised the design and simulation of robots with nanoscale dimensions, while the other embraced the idea of manipulation or assembly of nanoscale components using macroscopic instruments.

The first technical paper on nanomedical device design was published in 1998 by Robert A. Freitas Jr., J.D. titled "A Mechanical Artificial Red Cell: Exploratory Design in Medical Nanotechnology," which was the first book-length technical discussion of the medical applications of nanotechnology and medical nanorobotics[4].

As of now, a major part of the study in nanorobotics is being conducted in the field of biology and life sciences (i.e., bionanotechnology). Therefore, medical technology will be the first one to extract the advantages of this upcoming technology.

ARCHITECTURE OF NANOROBOTS[5]

In manufacturing nanorobots, the ambient conditions under which a nanorobot is operating like temperature range, pressure, density, electromagnetism (including light, and so on) are taken into consideration. (Medically, in future, these conditions may be the considered inside the veins of a human body where a nanorobot will carry out its functions). To meet these requirements, signal processing, data transmission, power supplies, and so on are carried out using nanotechnology, along with the help of VLSI (Very Large Scale Integration) design. Moreover, it has also been noticed that VLSI design using CMOS (Complementary Metal Oxide Semiconductor) by the means of ultraviolet lithography tends to attain the highest degree of accuracy.

Likewise, Very High Speed Integrated Circuit Hardware Description Language, also popularly known as VHDL, which is one of the most common integrated circuits (ICs) in the industry, can prove to be a great help in the implementation of designs and simulation of nanorobots to achieve even a higher level of accuracy and sensitivity.

Nanophotonic and nanotubes jointly may accelerate further the actual levels of resolution ranging from 248nm to 157nm devices in the assembly process of manufacturing of nanorobots in a CMOS industry.

Chemical Sensor

Nanowires[6] (Figure 1) have two quantum-controlled directions, in comparison to other low-dimensional systems, with one direction left unrestrained for the purpose of electrical conduction. As a result of this, nanowires can be used as nanoconductors for numerous nanodevice applications. Nanowires possess splendid properties for applications in various filed like that of excellent electrical, optical, mechanical, piezoelectrical, and field emission.

Nanowires, due to their spectacular properties, are capable of decreasing drastically the self-heating and thermal coupling for a CMOS when used as a suspended array in silicon circuits. Consequently, this provides maximum efficiency for applications regarding chemical changes, enabling new nanomedical applications. Therefore, more than half of the power supply in a nanorobot can be decreased by use of nanowires in their various circuits while manufacturing it.

N

Energy Supply

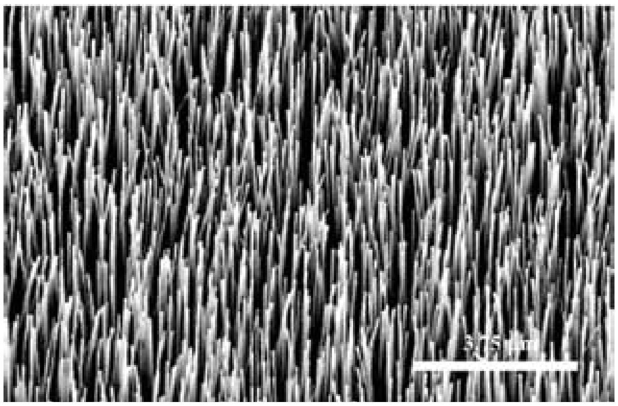
There can be a number of ways to supply energy to nanomachines from a human body itself. Inside a human body, temperature displacements can create potential differences, which can further be used as a voltage difference (Freitas, 1999).

The relation $E=hf$ can be well-used while using electromagnetic radiation, or light, where E is the energy generated by *photons,* depending on the frequency of light inside the body, f, encountering nanobots, and h is the Planck's constant (=6.626 x 10^{-34} Js). But these supplies of energy are not constant and may vary to a large extent. Thus, a wireless energy supply source may be used such as a microwave or a Radio Frequency wave generated from the patient's mobile phone to operate the transducers and *nanosensors* of nanobots efficiently inside a human body.

CURRENT DEVELOPMENTS IN NANOROBOTICS

Globally, several researches are being conducted in the field of nanorobotics in order to achieve the objective of synthesizing these miniature gadgets for human

Figure 1. SEM image of the zinc oxide (ZnO) nanowires grown by VLSI approach using Au catalyst (Courtesy: Liu, Wang, Kuo, Liang, and Chen (2007) "Recent Patents on Fabrication of nanowires")



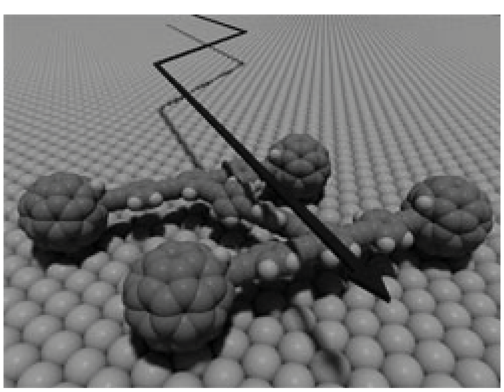
welfare. It is expected that within the next ten years, the development of nanobots will become comparable to the size of bacteria (Cavalcanti et al., 2004). Although nanorobots can prove to be a boon for the entire healthcare industry, the design of these nano miracles will be highly complex. Hence, the development of nanorobots in biomedical engineering can result in extremely complicated fabrication and control challenges. Also, a lot of research is currently being carried out all throughout the world using CMOS in VLSI design, with the help of VHDL in order to find a way to design and manufacture nanorobots. Quantum mechanics, thermal motions are also being considered in the design of nanobots.

Major development in the field of nanorobotics has been demonstrated by the Rice university scientists, which have constructed the world's smallest car—a single-molecule "nanocar" (Figure 2) that contains a chassis, axles, and four buckyball wheels.

According to Dr. James M. Tour, the Chao Professor of Chemistry, professor of mechanical engineering and materials science and professor of computer science, "*The synthesis and testing of nanocars and other molecular machines is providing critical insight in our investigations of bottom-up molecular manufacturing.*" [7]

The creation of the nanocar took over eight years. The entire nanocare[7] measures approximately 3–4 namometer, slightly wider than a DNA strand. It consists of a chassis and axles made of well-defined organic groups with pivoting suspension and freely rotating axles. The wheels are buckyballs, spheres of pure carbon containing 60 atoms apiece. This nanomachine is the first of its kind that actually functions like a car. It rolls on four wheels in direction perpendicular to its axles.

Figure 2. Nanocar (Courtesy: http://www.rpi.edu/~tsaiw/plans.htm)



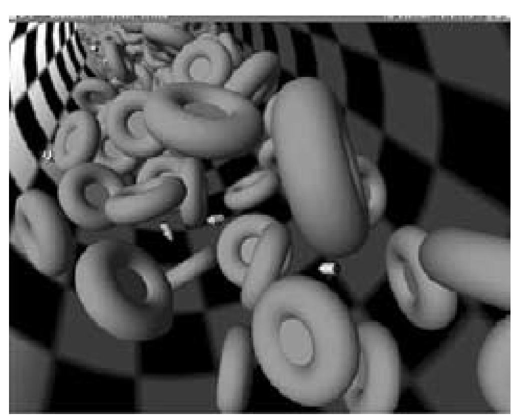
FUTURE APPLICATIONS OF NANOROBOTICS

In future, nanorobotics can prove to be extremely beneficial for the fields of information and communication, chemistry, environment, energy, and consumer goods, most of which will be reaped mainly by medicines and healthcare. However, there is still a lot to work that is to be done in order to meet the requirements of the synthesis of a nanodevice, which can be helpful to the entire healthcare industry, along with other areas of application. It might take a long time to realize all the theoretical concepts which are currently in the scene to manifest into reality.

By 2020, it is expected that nanobots will be used at a high level, and this technology may prove miraculous to the patients suffering from incurable diseases like diabetes (Figure 3). One may also be able to not only diagnose, but also control the BGL (Blood Glucose Level) by the use of nanorobots (Cavalcanti, Shirinzadeh, Freitas, & Kretly, 2007).

The two chief areas of applications of nanorobots, in future, may be categorised into Nanodiagnosis and Nanotherapeutics. Both these areas are briefly discussed in the following paragraphs:

Figure 3. Nanorobots in the diagnosis of glucose level in bloodstream (Courtesy: Cavalcanti, et al., "Medical Nanorobot Architecture Based on Nanobioelectronics")



- **Nanodiagonisis.** It is one of the most challenging aspects of nanorobotics. Continuous monitoring of patient's body condition can help in neurosurgery, detection of cancer at an early stage, and blood pressure control for cardiology problems. The same approach is quite useful in monitoring patients with diabetes. A RF wave transceiver like mobile phones can be of a great use in nanorobotic therapy, as these can be used for dual purposes including (a) a continuous source of energy for nanodevices, and (b) data reports can be transmitted to the mobile from nanodevices with a preinstalled software on the transceiver (Cavalcanti et al., 2007).•
- **Nanotherapeutics.** Cell herding machines are still a theoretical concept to stimulate rapid healing and tissue reconstruction along with cell repair machines to perform genetic surgery. These two areas of technology, if successful, will not only bring a revolution in the medical world, but also will prove to be a great benefit for the mankind. Nanorobotics will be able to diagnose, as well as the cure, the diseases quickly, efficiently, and effectively. Moreover, accurate and effective diagnosis of incurable diseases like cancer and HIV/AIDS at early stages will turn out to be a great relief for the coming generation from these deadly diseases.

CONCLUSION

Nanorobots can manipulate the objects at nanolevel or nanoscale. Nanorobotic instruments such as atomic force microscope can also perform manipulation at nanoscale. A lot of artificial nonbiological nanorobots are still in the developing phase, and will soon be a great benefit for the patients suffering from fatal diseases like Cancer, HIV/AIDS, and so on. Moreover, it will prove to be a milestone the medical world.

Nanorobotics has an enormous potential for phenomenal works in the fields of medicine and healthcare. In future, nanorobotics can help in the early and effective diagnosis of various diseases like cancer, HIV/AIDS, Alzheimer's, and so on.

The most important application of nanorobotics is in the field of medical technology. Currently, this technology is in its initial phase, and hence, more of theoretical concepts instead of practical are in scene as of now, in conjunction with a lot of research going on in this field worldwide. The theory of all these researches may be different but the goal is the same; that is, to promise the new generation a disease-free and healthy life with quick diagnosis and quick treatment, with the help of these miniature wizards called "nanorobots."

REFERENCES

Beem, H. (2006). Nanorobotics design, assembly and role in medicine. *Nanotechnology: From synthesis to self-assembly*. Retrieved February 11, 2008, from http://frontpage.okstate.edu/nanotech/Reports/Review%20Papers/FinalPaper%20Beem.pdf

Cavalcanti, A., & Freitas, R.A. (2002). Autonomous multi-robot sensor-based cooperation for nanomedicine. *Intl. J. Nonlinear Sciences and Numerical Simulation, 3*, 743-746.

Cavalcanti, A., & Freitas, R.A. (2005). Nanorobotics control design: A collective behavior approach for medicine. *IEEE Trans Nanobioscience, 4*(2), 133-140.

Cavalcanti, A., Freitas, R.A., & Kretly, L.C. (2005). Nanorobotics control design: A practical approach tutorial. *Robotics Today, 18*(4), 1-23.

Cavalcanti, A., Hogg, T., Shirinzadeh, B., & Liaw, H.C. (2006). Nanorobot communication techniques: A comprehensive tutorial. *IEEE ICARCV Int'l Conf. On Control, Automation, Robotics and Vision* (pp. 2371–2376).

Cavalcanti, A., Rosen, L., Shirinzadeh, B., & Rosenfeld, M. (2006). Nanorobot for treatment of patients with artery occlusion. *Springer Proceedings of Virtual Concept.* Retrieved February 11, 2008, from http://www.nanorobotdesign.com/papers/cardiology.pdf

Cavalcanti, A., Shirinzadeh, B., Freitas, R.A., & Kretly, L.C. (2007). Medical nanorobot architecture based on nanobioelectronics. In *Recent Patents on Nanotechnology 2007* (Vol. 1, No. 1, pp. 1–10). Bentham Science.

Freitas, R.A. (1999). Basic capabilities, landes, biosciences. *Nanomediceine*, *1*. Retrieved February 11, 2008, from http://www.nanomedicine.com/NMI.htm

Li, W.J., Xi, N., Fung, W.K., & Wong, T.S. (2004). Nanorobotics and nanomanipulation. *Encyclopedia of Nanoscience and Nanotechnology* (Vol. 7, pp. 351-365). American Science Publishers.

Liu, C., Wang, R., Kuo, C., Liang, Y., & Chen, W. (2007). Recent patents on fabrication of nanowires. In *Recent Patents on Nanotechnology 2007* (Vol .1, No. 11, pp. 11-20). Bentham Science.

Nain, A.S., Wong, J.C., Amon, C., & Sitti, M. (2006). Drawing suspended polymer micro/nanofibers using glass micropipettes. *Applied Physics Letters, 89*(18), 183105-183107.

Requicha, A.A.G. (1999). Nanorobotics. In S. Not (Ed.), *Handbook of Industrial Robotics*, (2nd Ed., pp. 199–210). John Wiley and Sons, New York.

Requicha, A.A.G. (2003). Nanorobots, NEMS and nanoassembly. In *Proceedings of the IEEE, Special issue on Nanoelectronics and Nanoprocessing* (Vol. 91, No.11, pp. 1922-1933).

Rosenberg, L. (2001). A review of nanomedicine. *New Surgery*. Retrieved February 11, 2008, from http://www.landesbioscience.com/journals/newsurgery/1.1/rosenberg.html

Seeman, N.C. (2005). Structural DNA nanotechnology: An overview. *Methods Mol Biol., 303*, 143-66.

Winstead, E.R. (2004). Robots in the body. *Genome News Network*. Retrieved February 11, 2008, from http://www.genomenewsnetwork.org/articles/2004/08/19/nanorobots.php

KEY TERMS

Blood Glucose Level (BGL): The concentration of glucose in the blood, measured in milligrams of glucose per 100 milliliters of blood.

Nanomachine: Devices, such as electronic circuits, from single atoms and molecules based on nanotechnology.

Nanoparticles: A microscopic particle whose size is measured in nanometres (nm). It is defined as a particle with at least one dimension <100nm.

Nanophotonic: The study of the behavior of light on the nanometer scale.

Nanoscale: Occurring on a scale of nanometers.

Very-Large-Scale Integration (VLSI): The process of creating integrated circuits by combining thousands of transistor-based circuits into a single chip.

VHSIC Hardware Description Language (VHDL): It is used as a design-entry language for field-programmable gate arrays and application-specific integrated circuits in electronic design automation of digital circuits.

ENDNOTES

1 http://en.wikipedia.org/wiki/Nanorobotics
2 http://www.zyvex.com/nanotech/feynman.html
3 http://en.wikipedia.org/wiki/K._Eric_Drexler
4 http://www.rfreitas.com/
5 http://www.bentham.org/nanotec/samples/nanotec1-1/Cavalcanti.pdf
6 http://www.bentham.org/nanotec/samples/nanotec1-1/Chuan-Pu%20Liu.pdf
7 http://www.media.rice.edu/media/NewsBot.asp?MODE=VIEW&ID=7850&SnID=1080666260

Networks of Action for Anti Retroviral Treatment Information Systems

Elaine Byrne
University of Pretoria, South Africa

Roy D. Johnson
University of Pretoria, South Africa

INTRODUCTION

The South African government has an impressive constitution and legislative framework that recognizes the right of its citizens to quality health care (Government of South Africa, 1996). In South Africa, approximately 80% of the population relies on state-provided health care. Health workers in the public health sector provide services at the formal health facilities and to the various outreach programs in the community (i.e., immunization drives). The effective management and delivery of these diverse services requires regular reporting of routine and exceptional information by health care workers. These workers spend a significant amount of time collecting, recording, storing, and transmitting data in various forms.

With the commencement of the Anti-Retroviral Treatment (ART) program in selected clinics throughout South Africa in 2003 (Department of Health, 2003), treating and supporting clients attending ART clinics places great pressure on the health staff, not only because of insufficient human resources and time, but also with the associated severe emotional strain. Pressure is escalating as the number of clients requesting ART is increasing daily (Stewart, Padarath & Bamford, 2004). An effective Information System (IS) is needed to manage this increase in clients as well as support a variety of reporting requirements.

A national survey in South Africa of health personnel, ambulatory and hospitalized patients, and health facilities substantiates that a weak patient IS (a) was an impediment to ensuring ongoing and correct treatment, (b) increased staff workloads, and (c) led to unnecessary duplication of effort and time. Additionally, Shisana, et al. (2002) argue that ensuring that a single electronic IS is in place to assist in treatment of patients is an essential yet often neglected aspect of the health system.

In 2005, the clinical director of the Batho Pele clinic[1] in the Gauteng province in South Africa requested the assistance of the Department of Informatics at the University of Pretoria in addressing their IS issues. This request fitted the department's research interests in health information systems (HIS), as well the broader research focus and commitment to provide outreach services to the community. Knowing the problems of commencing projects without having planned for sustainability and scalability, the HIS research group elected to use the "networks of action" concept to partner and collaborate with the various role players, institutions, and other ART entities. This process of developing interconnecting networks of human and nonhuman entities in South Africa and beyond its borders raised a number of opportunities, challenges, and tensions in initiating this project.

To provide a background to this process, the next section introduces the concept of "networks of action" and a brief description of the ART clinic. The following section develops the main focus of this chapter, which is the process of developing these networks. The last section suggests the necessity of developing networks of action as a future trend for sustainable IS.

BACKGROUND

Networks of Action

In addressing why so many action research efforts fail in the long term, Braa, Monteiro, and Sahay (2004) argue that the two major challenges in the development of a successful HIS are the interrelated factors of sustainability and scalability. Sustainability refers to making the IS work over time through the institutionalization of routines and the development of local learning processes. Scaling concerns the spreading of a working solution to other sites (Braa et al., 2004). However, scalability is not merely a technical problem but encompasses a sociotechnical network, comprised of people, technology, and processes within an institutional context and relates to a process of specifically what is being scaled and how it is being scaled (Sahay & Walsham, 2006).

As argued by Braa, et al. (2004), scalability requires local interventions to be part of or connected to broader networks in order for sustainability to occur. They argue that local action research interventions need to be conceptualized and approached as one element in a larger network of action in order to ensure sustainability. Sustainability cannot occur just through action at a local level, and scaling needs to occur through the creation of multiple interconnecting networks. A flexible and adaptive process, which accommodates planned and unplanned events, or as Giddens (1984) would say, anticipated and unanticipated consequences, needs to be adopted in order for scaling to occur successfully.

Building on the Scandinavian-based action research's recognition of the need to perform action as part of a network rather than as a sole local venture (Chisholm & Elden, 1993; Elden & Chisholm, 1993), Braa, et al. (2004) argue that the need to develop an institutionalized and sustainable system is not a luxury but a necessity and needs to be part of larger interventions. Networking enables the sharing of experience, knowledge, and technology and thereby scaling the learning process. A key concept from this learning process is an alternative approach to action research; namely, the development of networks of action. These are characterized as:

1. "Abandoning singular, one-site action research projects in favor of a network of sites
2. Generating local, self-sufficient learning processes together with working mechanisms for the distribution of appropriately formatted experiences across sites in the form of vertical and horizontal flows
3. Nurturing a robust, heterogeneous collection of actors likely to pursue distinct yet sufficiently similar agendas
4. Aligning interventions with the surrounding configurations of existing institutions, competing projects, and effort, as well as everyday practices" (Braa et al., 2004, p. 359)

The development of networks of action is pivotal to addressing the challenges of sustainability and scalability and is especially important to the project described in this chapter, the Anti-Retroviral Treatment Information System (ARTIS) Project.

Background of ART Program

The Batho Pele ART clinic was recently established (2006) and is housed in a district hospital in the province of Gauteng. The clinic is part of the government's plan to improve the delivery of HIV/AIDS-related services (Department of Health, 2003).

Principally, there are three phases in processing a patient at this ART clinic:

- **Phase 1. Making an initial appointment.** This is based on the patient having a CD4 count <200 cells/mm3 or a WHO Stage IV disease (Department of Health, 2004). The patient has to be referred to the clinic from another clinic or medical practitioner.
- **Phase 2. Preprescription of ART.** Depending on the health and emotional status of the patient, acceptance for ART can occur, starting with the third visit or later.
- **Phase 3. ART prescription and follow-up.** Once accepted onto ART, there are regular return visits scheduled for pill counting, patient assessment, and prescription renewal.

Each visit requires the patient to check in with the ART administrator and register with the hospital administrator. Once registered, the patient is seen on each visit by nurses for tests (i.e., blood, weight, urine, blood pressure) and by the doctor, who reviews and analyzes the test results and prescribes the treatment regimen. The pharmacist dispenses the antiretroviral drugs. Consultations with the counselor, social worker, and dietician occur as the need arises (De Freitas, 2005).

Although there are official workflow processes for an ART clinic, these workflow processes are not adhered to stringently for a variety of reasons. Staff members are allocated to particular positions with clearly defined roles and duties, but due principally to shortages of staff, little segregation of duties occurs. When there is a shortage of staff or a staff member is experiencing a heavy workload, it is common for staff members to assist in completing the work tasks of their colleagues. For example, the administrator will assist the data entry operator when few patients are waiting at the reception. This implies that staff members are performing duties for which they have not received an adequate orientation, professional training, or supervised practice. This causes problems on many levels, such as security for accessing data and the quality of data entered incorrectly, incompletely, or duplicated (De Freitas, 2005).

Based on this background of the meaning of networks of action and the situation at the ART clinic, the next section describes our experience of developing these networks.

ESTABLISHING NETWORKS OF ACTION

From the preliminary review of the paper-based IS currently in place at the District Hospital ART clinic (De Freitas & Byrne, 2006), the IS does not meet the requirements of the staff and hence does not support the effective and efficient delivery of services to ART clients.

The IS at the ART clinic is a paper-based system and a partial computer-based system using an Excel spreadsheet. The paper-based system consists of a patient file, which contains all the necessary health-related documents on the patient's health condition and is stored in a filing room. The longer a patient stays on treatment, the bulkier and more cumbersome the documents become. The process of reviewing data in the paper file then becomes unwieldy, frustrating, and time consuming. The computer-based part of the system is an Excel spreadsheet originally designed at another ART clinic. These Excel spreadsheets contain basic demographic data, CD counts, and appointments.

In these initial stages of design and development, a conscious effort was made to establish networks of action. This process has been time consuming, frustrating, and confusing. However, the focus on sustainability overcomes these challenges. The various actors involved to date in the process of developing our network of actions are as follows:

- **Alignment of interests and focus.** In 2005, the Informatics Department established an HIS research group. The team meets regularly, and HIS has become one of three research focus areas for the department.
- **Memorandum of understanding.** A memorandum of understanding, which clearly outlines the role and responsibilities of the clinic and the university, has been established between the ART clinic and the Department of Informatics. The ART clinic agrees to give access at prearranged times; participate in IS requirements analysis, design, and implementation; and provide staff training on the system. The Department of Informatics agrees to perform an IS requirements analysis and develop a system; support the IS; assist in training health staff and students; and conduct research in IS design, development, and training.
- **User requirements.** A preliminary review of the IS requirements was obtained in 2005. In early 2006, an initial group meeting was held with all the clinic staff on the user requirements. However, at the following meeting, none of the clinic staff showed up. Later, they were evasive about the lack of support from the clinic staff but showed incredible interest in the documentation and insisted on keeping it.
- **Building upon networks.** The department has relationships with HIS research networks, such as Informatics Development for Health in Africa (INDEHELA) and Building Europe Africa collaborative Network for applying IS (BEANISH). INDEHELA is a Finnish initiative for collaborative research between Finland and Africa (Nigeria, Mozambique, and South Africa). Part of their research focuses on sustainability and contextual HIS design. BEANISH is a collaborative HIS initiative whose overall goal is to build networks of researchers and development between and within countries in Africa and Europe by studying practical applications of IS technology in the health sector.
- **Linking existing ART modules or systems.** One of the first steps of our project was a review of the existing ART modules or open source patient-based HIS. The Care2x hospital management system, the OpenMRS data model, the District Health Information System (DHIS) ver1.4 and 2.0, the Ethiopian ART module, and the Excel spreadsheet from another hospital have been explored to varying degrees. An ART module prototype has been developed and the possibility of exporting aggregated data to the DHIS, which is currently the official district HIS in South Africa

(HISP, 2006), and the potential of integrating as a module into Care2x is being investigated.

The most intensive review was a workshop in October 2006 on the refactoring of an existing ART module running in clinics of Ethiopia. The ART workshop was facilitated by the Open Source Centre from the Council of Scientific and Industrial Research in South Africa, and the refactoring process continued at the center the following week. Representatives from HIS networks in Ethiopia, Norway, India, and Vietnam attended the workshop. The development of the ART prototype took place during December 2006 and January 2007. Deployment of the system is still underway at the time of this writing.

This review of the process of developing the networks of action has implications for future IS development projects and are now discussed.

FUTURE TRENDS

From the review of the process of the development of the networks of action, it is apparent that this is a time-consuming process. However, developing such networks should be viewed as a necessity for the long-term sustainability of the systems. There is a number of implications of adopting such an approach, which surfaced within our process.

- **Time constraints.** There is a constant battle to deliver health services to implement an ART solution quickly, yet the need is not to reinvent the wheel or develop parallel systems. An agile and staged approach to systems development is appropriate, but the scope still needs to be delimited.
- **Various approaches.** Transferring tools and techniques from one sector to another may not be appropriate. Gaining user requirements from health practitioners may be more fruitfully done through participatory prototyping as illustrated in the development of DHIS rather than gaining all user requirements upfront.
- **Partnerships.** Within the health sector, there is a need to build relationships with the clinic staff, middle and top management, and across research groups, universities, clinics, hospitals, government departments, and NGOs. Cultural differences and different work practices across cultures also pose a challenge.
- **Understanding existing artifacts.** Given the number of patient-based HIS and ART modules, the review of existing artefacts could be paralyzing. This, however, raises opportunities for pooling resources (expertise and financial), which would otherwise not be possible.
- **Institutional building.** The institution in which the research is to be conducted needs to be clearly defined, establish institutional backing, and have a long-term commitment.

Insufficient attention to the process of establishing networks can severely impact the sustainability and scalability of the IS. However, the process of establishing these networks is not generally described in action research projects or in IS development initiatives. From our experience, this process has been time-consuming but needs to form a substantial part of any IS that wishes to be sustainable and scalable.

CONCLUSION

Within the IS discipline, there has been a movement away from the previously dominant technical orientation of systems developers, stemming primarily from a computer science tradition to include and actively recognize the needs, aspirations, and expertise of users and to understand IS as social processes (Cornford & Smithson, 1996; Roode, 2003). However, there is little focus on how to develop the networks of action within this sociotechnical process. Missing this vital process in terms of planning and allocating resources can have a negative impact on IS projects. This stage of IS design and development should receive more attention in project management and IS development methodology.

REFERENCES

Braa, J., Monteiro, E., & Sahay, S. (2004). Networks of actions: Sustainable health information systems across developing countries. *Management Information Systems Quarterly, 28*(3), 337–362.

Chisholm, R.F., & Elden, M. (1993). Features of emerging action research. *Human Relations, 46*(2), 275–298.

Cornford, T., & Smithson, S. (1996). *Project research in information systems: A student's guide*. London: Macmillan.

De Freitas, M.R. (2005). *Quality and use of data for decision support: A case study of the antiretroviral clinic*. Pretoria, South Africa: Department of Informatics, University of Pretoria.

De Freitas, M.R., & Byrne, E. (2006). *Activity theory as an analytical tool: A case study of IS development for an anti-retroviral treatment clinic in South Africa*. Proceedings of the South African Institute of Computer Scientists and Information Technologists (SAICSIT) conference, South Africa, 90–99.

Department of Health, South Africa. (2003). *The operational plan for comprehensive HIV/AIDS treatment.* Retrieved September 2006, from http://info.gov.za/issues/hiv/careplan.htm

Department of Health, South Africa. (2004). *National antiretroviral treatment guidelines*. Retrieved June 2006, from www.doh.gov.za/docs/factsheets/guidelines/artguidelines04/intro.pdf

Elden, M., & Chisholm, R.F. (1993). Emerging varieties of action research: Introduction to the special issue. *Human Relations, 45*(2), 121–142.

Giddens, A. (1984). *The constitution of society: Outline of the theory of structuration*. Cambridge: Polity.

Government of South Africa. (1996). *Constitution of the Republic of South Africa*. Act 108 of 1996 as adopted on 8 May 1996 and amended on 11 October 1996 by the Constitutional Assembly. Retrieved October 2006, from http://www.polity.org.za/html/govdocs/constitution/saconst.html?rebookmark=1

HISP. (2006). *History of HISP-SA*. Retrieved October 2006, from http://www.hisp.org/

Roode, D. (2003). Information systems research: A matter of choice? *South African Computer Journal, 30*, 1–2.

Sahay, S., & Walsham, G. (2006). Scaling of health information systems in India: Challenges and approaches. *Journal of Information Technology for Development, 12*(3), 185–200.

Shisana, O., et al. (2002). *The impact of HIV/AIDS on the health sector*. National survey of health personnel, ambulatory and hospitalised patients and health facilities. A collaborative effort of the Human Sciences Research Council (HSRC), the Medical University of Southern Africa (MEDUNSA) and the Medical Research Council (MRC).

Stewart, R., Padarath, A., & Bamford, L. (2004). *Providing antiretroviral treatment in Southern Africa. A Literature Review*. South Africa: Health Systems Trust.

KEY TERMS

Action Research: A collaborative cyclical inquiry process that involves diagnosing a situation, taking action, and reflecting on the outcome before recommencing the process.

ART: The treatment of people infected with HIV/AIDS by incorporating antiretroviral drugs, counseling, and guidance on how to live a healthier life.

Developing Countries: Countries generally found in the southern hemisphere and characterized by low levels of human development.

Health Information Systems: Sociotechnical combinations of human and nonhuman artefacts for coordination and communication work, or social activities in the health sector.

Networks of Action: A term used in action research that involves the cultivation and alignment of groups of people, institutions, sites, and artefacts.

Scaling: The spreading of working sociotechnical solutions to other sites, including people, technology, and processes.

Sustainability: Involves a long-term working solution.

New Ensemble Learning Approaches for Microarray Data Classification

Ching-Wei Wang
University of Lincoln, UK

INTRODUCTION

A reliable and precise classification of tumors is essential for successful treatment of cancer. Microarray technologies allow the rapid and comprehensive assessment of the transcriptional activity of a cell, leading to a more comprehensive understanding of the molecular variations among tumors and, hence, to a finer informative classification. However, the major challenge in using this technology is the analysis of its massive data output, which requires powerful computational means for interpretation. Hence, the ability to interpret the information in gene expression data becomes a critical issue in Genetics.

One of the most active areas of research in supervised machine learning has been to study methods for constructing good ensembles of classifiers. The main discovery is that the ensemble classifier often performs much better than single classifiers that make them up. Recent researches (Dettling, 2004, Tan & Gilbert, 2003) have confirmed the utility of ensemble machine learning algorithms for gene expression analysis. The motivation of this work is to investigate a suitable machine learning algorithm for classification and prediction on gene expression data.

The research starts with analyzing the behavior and weaknesses of three popular ensemble machine learning methods—Bagging, Boosting, and Arcing—followed by presentation of a new ensemble machine learning algorithm. The proposed method is evaluated with the existing ensemble machine learning algorithms over 12 gene expression datasets (Alon et al., 1999; Armstrong et al., 2002; Ash et al., 2000; Catherine et al., 2003; Dinesh et al., 2002; Gavin et al., 2002; Golub et al., 1999; Scott et al., 2002; van 't Veer et al., 2002; Yeoh et al., 2002; Zembutsu et al., 2002). The experimental results show that the proposed algorithm greatly outperforms existing methods, achieving high accuracy in classification.

The outline of this chapter is as follows: Ensemble machine learning approach and three popular ensembles (i.e., Bagging, Boosting, and Arcing) are introduced first in the Background section; second, the analyses on existing ensembles, details of the proposed algorithm, and experimental results are presented in Method section, followed by discussions on the future trends and conclusion.

BACKGROUND

Ensemble methods are learning algorithms that construct a set of base classifiers and then classify new data points by taking a vote of their predictions. The spirit in ensemble machine learning is to combine a number of rough "rules-of-thumb" into a more accurate aggregate class prediction rule. The learning procedure for ensemble algorithms can be divided into the following two parts. The first stage is constructing base classifiers/base models. The main tasks of this division are (1) data processing: prepare the input training data for building base classifiers by perturbing the original training data; and (2) base classifier constructions: build base classifiers on the perturbed data with a learning algorithm as the base learner. In this project, the C4.5 decision tree algorithm (Quinlan, 1996) is employed as the base learner. The second stage is voting, which combines the base models built in the previous stage into the final ensemble model. There are various kinds of voting systems. Two main voting systems are generally utilized; namely, weighted voting and unweighted voting. In the weighted voting system, each base classifier holds different voting power. On the other hand, in the unweighted system, an individual base classifier has equal weight, and the winner is the one with the most number of votes. Figure 1 illustrates the structure of the ensemble machine learning approach.

There are three types of generally adopted ensembles (i.e., Bagging, Boosting, and Arcing). Bagging algorithm, which is introduced by Breiman (1996), constructs base classifiers with inputs generated by the bootstrapping technique. The construction process of every base classifier is independent of each other. It perturbs the training set repeatedly to generate multiple predictors and combines these base classifiers by simple voting (classification) or averaging (regression) in order to obtain an aggregated predictor. The multiple input data for building base classifiers is formed by bootstrapping replicates of the original learning data.

Boosting was introduced by Schapire (1990) as a method to enhance the performance of a weak learning algorithm. Freund and Schapire (1996) proposed an algorithm called AdaBoost. There are lots of varieties of Boosting algorithms, and AdaBoostM1 is chosen as the Boosting method used in this project. Boosting adaptively reweights the training set in a way based on an error rate of the previous base classifier. The Boosting algorithm improves its behavior in reflection to the latest faults it makes. Moreover, if the error rate of a base classifier is greater than 0.5 or equal to 0, the sequential construction of base classifiers stops.

The framework of Arcing introduced by Breiman (1998) is similar to the one employed in Boosting. They both proceed in sequential steps. The major difference between Arcing and Boosting is that Arcing improves its behavior based on the accumulation of its faults in history. It examines all previous base classifiers' faults for construction of a new base classifier, while Boosting only checks the previous one base classifier. Apart from this, Arcing adopts unweighted voting system, whereas Boosting uses weighted voting. In addition, unlike Boosting, no checking procedure exists through the constructions of base classifiers.

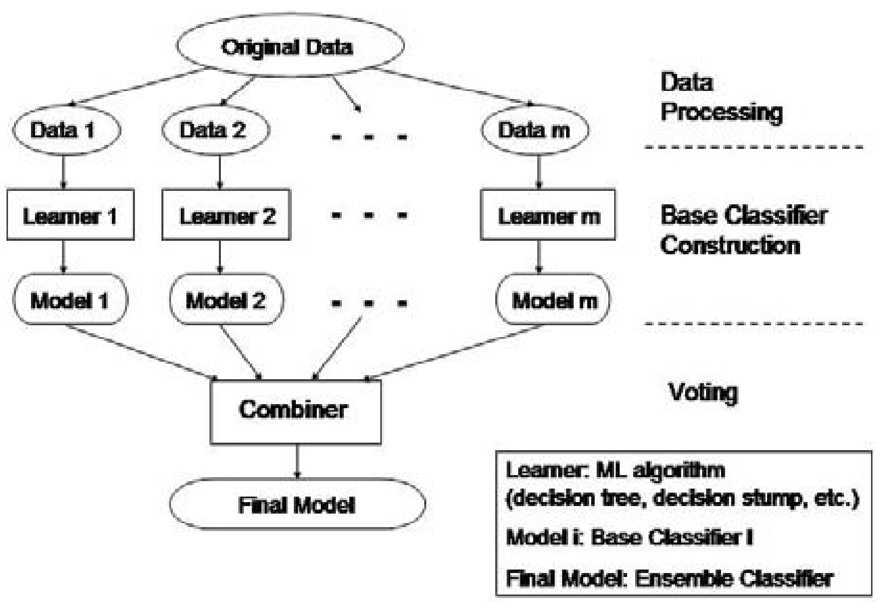
METHOD

In this section, the behavior and weaknesses of the existing ensembles are first analyzed, followed by presentation of the design and detailed algorithm of the proposed approach.

Behavior and Weaknesses of Existing Ensembles

1. **Boosting.** The accuracy of Boosting models will remain the same after specific numbers of base models are established, due to the checking mechanism following each construction of base classifiers. The specific criterion in Boosting stops further construction of base classifiers while its error rate is equal to 0 or greater than 0.5 (see Figure 2). Therefore, if the sequential construction halts after building six base classifiers, the same result will be obtained on evaluating over Boosting

Figure 1. Ensemble machine learning algorithm



models with any number greater than six, because fundamentally, these models are all identical. In other words, as long as the error rate of the first base classifier is equal to 0 or greater than 0.5, no matter how many base models specified, the whole construction terminates, and consequently, the entire ensemble model will be composed of exactly one base model.

This is an extreme case in Boosting, but it does happen very often in our experiments. As gene expression data consist of large amounts of genes, it helps learning methods to generate a more precise classifier that fits exactly on the training set. To sum up, due to the nature of gene expression data, the checking criterion seriously influences the diversity of boosted models by forbidding further construction of base models.

2. **Arcing.** The performance of Arcing models may deteriorate while more base classifiers are constructed. Unlike Boosting, without the checking condition interrupting, the other unwelcome situation may occur. The extreme undesirable result is an arced model with all identical base models. Without the checking criterion, Arcing may keep on producing the same base models. For example, once Arcing develops a base classifier that precisely fits the training data, there will be no misclassification value to be added in, and all misclassification values will remain the same. Hence, the instances' weights will stay the same as well. With the same data, the same instances' weights and the same base learner, the Arcing algorithm will produce exactly the same base classifier. As a result, due to the low diversity issue of base models, ensembles with self-optimization learning style may suffer from an over-fitting issue.
3. **Bagging.** Different from Arcing or Boosting, Bagging generates its base models by chance, because the constructions of the base models are independent of each other.

Design: Framework

The advantage of Boosting and Arcing is to refine their behavior based on previous learning experience. However, over-fitting the issue decreases the quality of these ensemble models. Hence, a data-perturbing technique is created for constructions of base classifiers in order to produce base models in higher diversity. Particularly, in construction of each base model, Boosting refines itself based on errors by the previous model, whereas Arcing learns based on errors by all previous models. It is found that the Boosting structure obtains higher accuracy (see Table I). The reason suggested here is that it is not necessary to pay attention to errors of all previous models since the errors have been corrected through construction of each base model. Figure 3 illustrates the overview of the proposed ensemble machine learning approach, which

Figure 2. Boosting: construction of base classifiers

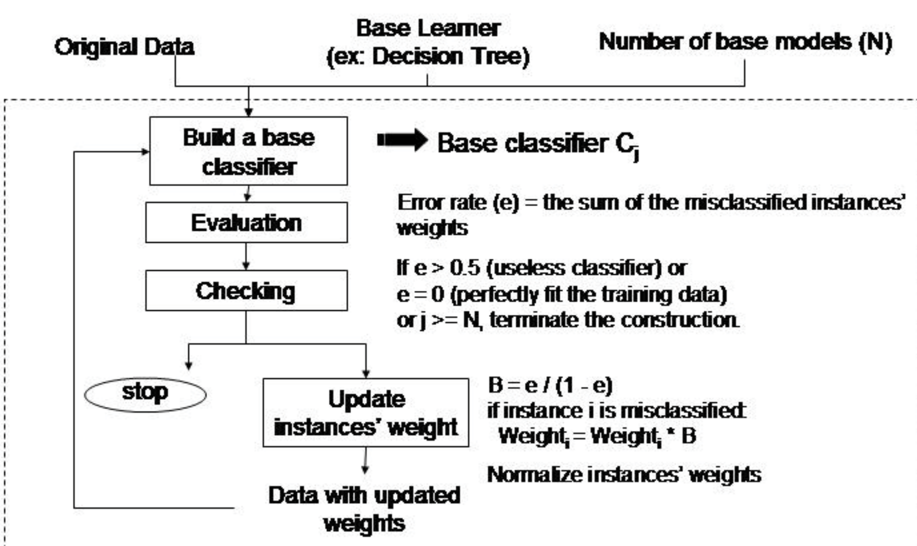
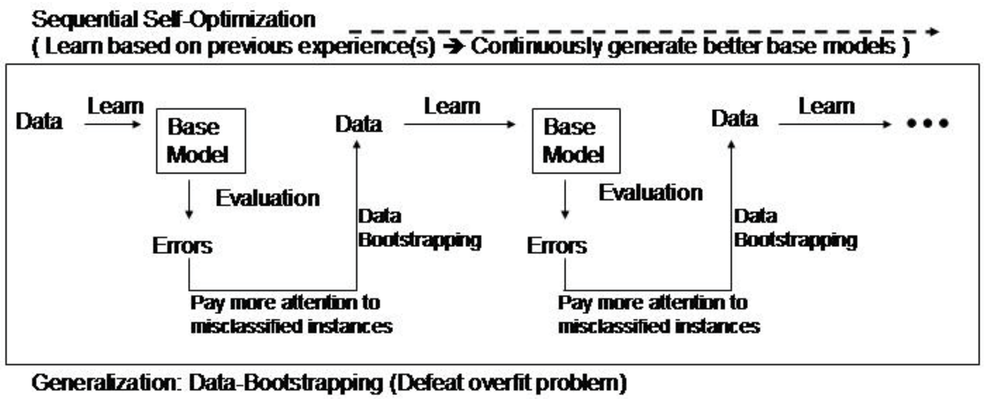


Figure 3. Behavior of the new ensemble machine learning approach



includes the sequential self-optimization structure and data-perturbing technique.

Design: Data Bootstrapping

The bootstrapping method is to generate a new dataset by sampling from the given dataset. Therefore, some instances may be selected several times, whereas others may be ignored. For every bootstrapping, a list of random floating numbers will be generated. The number of the random values is exactly the same as the size of input data set. These floating numbers are then processed into a list of accumulated probabilities, and the maximum value of this distribution is adjusted to the summation of instances' weights. Equation 1 presents the computation of each accumulated probability:

$$\mathrm{P(a)} = [\sum_{i=1}^{a} \mathrm{Random(i)} / \sum_{j=1}^{n} \mathrm{P(j)}] \cdot \sum_{k=1}^{n} W(k) \qquad (1)$$

Afterwards, the P(a) value is compared to the accumulated weight value ΣW(b).

$$\sum W(b) = \sum_{i=1}^{b} Weight(i) \qquad (2)$$

If the P(a) < ΣW(b), then add instance b into the result bootstrapped dataset, and the next comparison will be made between P(a+1) and ΣW(b). Otherwise, do not select instance b, and the next comparison will be made between P(a) and ΣW(b+1). Thus, one instance may be selected many times, and the higher its weight is, the greater the accumulated weight's value becomes, contributing to a bigger chance to be selected. The algorithm is devised in Figure 4.

Design: Main Algorithm

The proposed method is composed of the sequential self-optimization structure from Boosting and the data-bootstrapping technique. Figure 5 presents the details of the main algorithm.

Experimental Results

Cross-validation is deemed an objective and commonly used tool in model selection. Especially, cross-validation is markedly superior for datasets with a small number of samples, which exactly matches the gene expression data case; this fact is demonstrated in (Goutte, 1997). In this paper, tenfold cross validation is utilized for evaluation. Moreover, the experiments are conducted in a linux-based cluster (ScotGrid). Twelve gene expression datasets are obtained from published research works (Alon et al., 1999; Armstrong et al., 2002; Ash et al., 2000; Catherine et al., 2003; Dinesh et al., 2002; Gavin et al., 2002; Golub et al., 1999; Scott et al., 2002; van 't Veer et al., 2002; Yeoh et al., 2002; Zembutsu et al., 2002), and the C4.5 decision tree algorithm (Quinlan, 1996) is employed as the base learner. Three existing approaches (i.e., arcing, boosting, and bagging) and two new proposed methods (i.e., (New1) Data Bootstrapping + Arcing Skeleton and (New2) Data Bootstrapping + Boosting Skeleton) are tested.

Table 1 presents the percentage of accuracy in classification over 12 gene expression datasets by the

Figure 4. Data Bootstrapping Algorithm

Input: A dataset with n instances
Output: A bootstrapped dataset with the same size of the input dataset
Assumption:
Each instance has a weight, representing the relative level of importance in the dataset.
Process:
1. Generate n random numbers: R_1 to R_n.
2. Compute n accumulated probabilities: $P_i = \sum R_i$, where i =1 to n.
3. Normalize probabilities' values: $P_m = [P_m / \sum P_i] * \sum W_i$, where m, i = 1 to n
**Initialize two indices for iteration through two lists.
4. k = the index of the first probability in probabilities' list
5. x = the index of the first instance's weight in weights' list
** Iterate through lists of probabilities and instances' weights
6. Check if k <= n & x <= n. If both true, go to step 6.1
    6.1 $sumW(x) = \sum W_i$, where i = 1 to x
    6.2 Check if k <= n & P_k < sumW(x). if both true, go to step 6.2.1
        6.2.1 add instance x to the output dataset
        6.2.2 $W_k = 1$
        6.2.3 plus 1 to k.
    6.3 plus 1 to x.
7. Terminate with the output dataset

Figure 5. New ensemble machine learning algorithm

Inputs:
1. A training set T < X, Y >, where T: $\{<x_1, y_1>, ..., <x_n, y_n> \mid x \in X, y \in Y\}$.
    X: a set of instances: $\{x \mid x = <a_1, ..., a_q>\}$, where a_i = attribute value, q = number of attributes.
    Y: a set of classes, where y = 1 to z
2. Number of base classifiers R
3. The limit value of bootstrap times L
4. Base Learner / Inducer: C4.5 decision tree is used in this paper.

Output:
The Final model: Function C*
Steps:
1.Initialise instances' weights: For i = 1 to n, Weight: $W_1(i) = 1/n$
2.Generate a copy of the training data: S
  (T is for evaluating base classifiers. S is for building base classifiers.)
3.Repeat 3.1 to 3.9 R times
    1.Bootstrap S dataset from previous round S dataset
    2.Build a new Classifier $C_r(X)$ using weighted S dataset (X, W_r) by base learner.
    3.Evaluate $C_r(X)$ with T. Error rate e = Sum of the weights of the misclassified instances by $C_r(X)$
    4.Check if e = 0 & the number of data bootstrap ≤ L. If both true, go back to 3.1 to do the bootstrap.
    5.If e > 0.5 or e = 0, go to step 4.
    6.$B_r = e / (1 - e)$
    7. Set sum of instance weights for next round to 0: $SumW_{r+1} = 0$
    8. Update instances' weights: For i = 1 to n, check if Cr misclassifies instance i.
        1.If true, $W_{r+1}(i) = W_r(i) * B_r$. Otherwise, $W_{r+1}(i) = W_r(i)$
        2.$SumW_{r+1} = SumW_{r+1} + W_{r+1}(i)$
    9. Normalize instances' weights: For i = 1 to n, $W_{r+1}(i) = W_{r+1}(i) / SumW_{r+1}$
4.Produce ***Function C*(instance)*** by Voting

continued on following page

Figure 5. continued

Auxiliary algorithm: Voting

Function C*(instance)
Input: instance
Output: predicted result y_i
Steps:
1. Initialise votes of classes to 0: For i = 1 to z, V(i) = 0
2. For j = 1 to R, V(i) = V(i) + log (1 / B_j)
(i: class index from base model $_j$ in classifying the input instance.)
3. a = Argmax V(i), where i = 1 to z
4. Terminate with the predicted result y_a

five ensembles. The experimental results show that the proposed ensemble classifiers stably achieve high accuracy over all 12 gene expression datasets in comparison to existing ensembles. Particularly, the combination of data bootstrapping and boosting skeleton (New2) performs best in classifying gene expression data among five models.

FUTURE TRENDS

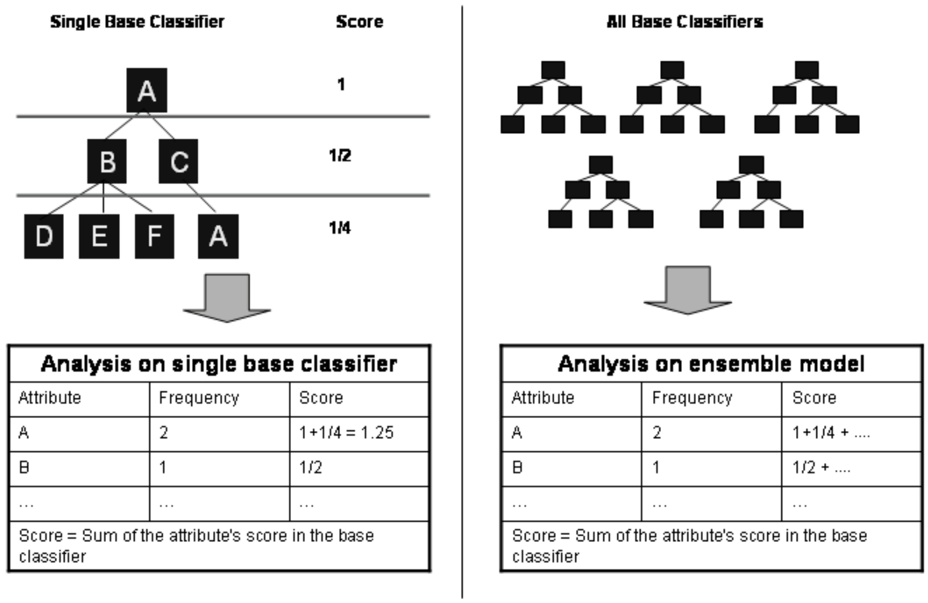
Obtaining patterns and rules with high accuracy in classification and prediction, a further investigation is to extract information about sets of influential attributes/genes from the resulting patterns and rules, allowing advanced study in biological meaning of genes and drug discovery.

The aim of discovery influential genes in microarray data analysist is to explore the relative importance of attributes within the ensemble model. Considering the patterns and rules learned, the most important criterion in classification will be placed as the first priority. As the decision tree algorithm is employed as the base learner in this study, the attributes in the upper tier of a tree are more meaningful in the classifying procedure. Hence, to obtain a comprehensive view on the influential genes/attributes in classification of microarrays, a design is proposed here to exhibit the attributes' precedence among all base models. The design scheme first assigns a weight for each tree level, and the attributes gain

Table 1. Prediction performance by 10-fold cross validation

Dataset	Arcing	Boosting	Bagging	New1	New2
AML ALL	88.9	91.7	94.4	100	100
Brain Cancer	84	82	84	94	94
Breast Cancer	80.41	85.57	90.72	94.85	95.88
CNS	78.33	90	88.33	93.33	95
Colon Tumor	69.35	80.65	79.03	79.03	83.87
Lung Cancer	97.24	97.24	97.79	98.34	99.45
Prostate Outcome	87.5	90.44	94.12	94.85	97.06
Prostate Tumor	66.67	76.19	61.9	95.24	100
DLBCL Outcome	84.48	93.1	98.28	100	100
DLBCL Tumor	85.71	94.81	92.21	97.4	98.7
ALL MLL AML	91.67	91.67	91.67	93.06	98.61
7 Subtypes	80.12	92.66	91.44	89.3	93.88

Figure 6. Computation on decision power of attributes to discover influential genes



points related to the weight in the particular tree level. Afterwards, the score of the individual attribute in each base model is associated to the voting power of the base model, which is multiplied by log ((1- error rate) / error rate). By summing up an attribute's points gained in all base models, its total score can then be obtained. The design scheme is presented in Figure 6.

CONCLUSION

Accompanied with the sequential self-optimization structure, the entire ensemble model is inclined to refine its behavior in a correct direction. The experimental results on 12 gene expression datasets prove that the proposed ensemble classifiers are largely upgraded with the data-bootstrapping technique. The combination of data bootstrapping and boosting skeleton (New2) is able to obtain high accuracy stably, and hence, it is recommended for gene expression data analysis.

REFERENCES

Alon, U., et al. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Cell Biology, 96*, 6745–6750.

Armstrong, S.A., et al. (2002). MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics, 30*, 41–47.

Ash, A.A., et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature, 403*, 503–511.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 26*(2), 123–140.

Breiman, L. (1998). Arcing classifiers. *Annals of Statistics, 26*(3), 801–849.

Catherine, L.N., et al. (2003). Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Research, 63*(7), 1602–1607.

Dettling, M. (2004). BagBoosting for tumor classification with gene expression data. *Bioinformatics, 20*(18), 3583–3593.

Dinesh, S., et al. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell, 1*, 203–209.

Freund, Y., & Schapire, R. (1996). *Experiments with a new boosting algorithm*. Proceedings of the Thirteenth International Conference on Machine Learning, 148–156.

Gavin, J.G., et al. (2002). Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Research, 62*, 4963–4967.

Golub, T.R., et al. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science, 286*, 531–537.

Goutte, C. (1997). Note on free lunches and cross-validation. *Neural Computation, 9*, 1245–1249.

Quinlan, J.R. (1996). *Bagging, boosting, and C4.5*. Proceedings of the Thirteenth National Conference on Artificial Intelligence, 725–730.

Schapire, R. (1990). The strength of weak learnability. *Machine Learning, 5*(2), 197–227.

ScotGrid. Retrieved from http://www.scotgrid.ac.uk/

Scott, L.P, et al. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature, 415*, 436–442.

Tan, A.C., & Gilbert, D. (2003). Ensemble machine learning on gene expression data for cancer classification. *Applied BioInformatics, 2*, 75–83.

van 't Veer, L.J., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature, 415*(6871), 484–485.

Yeoh, E.J., et al. (2002). Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell, 1*(2), 133–143.

Zembutsu, H., & et al. (2002). Genome-wide cDNA microarray screening to correlate gene expression profiles with sensitivity of 85 human cancer xenografts to anticancer drugs. *Cancer Research, 62*(2), 518–527.

KEY TERMS

Attributes: A property of an instance that may be used to determine its classification.

Base Learner: The algorithm used for building base classifiers (e.g, decision tree).

Base Model: The model generated by the base learner. A base classifier.

Classifier: A set of patterns and rules to assign a class to new examples.

Gene Expression Data: A typical gene expression array is made up of x columns of sample tissues, whereas each sample tissue contains y number of genes values. Every sample has its own individual class label, which represents the condition it belongs to.

Machine Learning (Mitchell, 1997): A computer system is said to learn from some experience E with respect to some class of tasks T and performance measure P, if it improves its performance (as measured by P) at tasks in T after passing the experience E.

Supervised Machine Learning: Machine learning algorithms require training datasets with prior knowledge of class labels.

New Perspectives in Rheoencephalography

Juan J. Perez
Polytechnic University of Valencia, Spain

Enrique Guijarro
Polytechnic University of Valencia, Spain

Pedro Ortiz
Consorcio Hospital General Universitario de Valencia, Spain

José M. Pons
Consorcio Hospital General Universitario de Valencia, Spain

INTRODUCTION

One of the most important advances in biomedical engineering has been the ability to inspect inside the body without opening it. In this sense, rheoencephalography (REG) is an electromedical technique used to assess the cerebral blood flow (CBF) by noninvasive electrical impedance methods, using electrodes attached to the scalp surface. This technique was first proposed by Polzer and Schuhfried (1950), and emerged as an extrapolation of impedance plethysmography applied to the head.

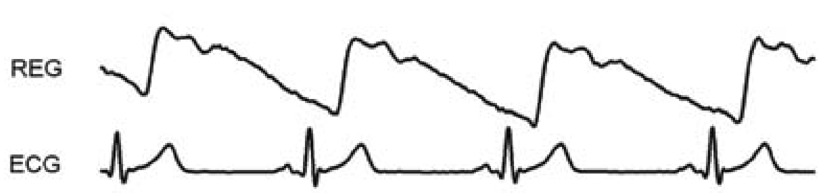
An electric current flowing through a biological tissue causes a potential difference between any pair of electrodes that can be measured. This potential difference depends on the amplitude of the injected current, the shape of the conductor, the arrangement of the electrodes, and the electrical characteristics of the tissue. For instance, the electrical conductivity of the lung tissue is much lower than that of the cerebrospinal fluid (CSF), since alveolar sacs are nonconductive. Furthermore, the electrical conductivity depends on the frequency of the electric current, the orientation of the tissue fibers relative to the current flow, and the amount of extracellular fluid that surrounds the cells. For example, electrical conductivity is higher in the blood than in most tissues, since plasma acts as a true-highway for ions (Malmivuo & Plonsey, 1995).

he ejected blood fills the arterial vascular bed, and a small and abrupt change in the electrical impedance of the tissues occurs, due to two reasons: the cross-sectional area of the tissue increases, and tissue effective conductivity rises. This is the physical principle of impedance plethysmography, and explains why the electrical impedance of any part of the body throbs synchronised to the heartbeat some tens or hundreds of milliohms around a basal impedance value.

BACKGROUND

By injecting an electric current through a pair of electrodes attached to the scalp surface, Polzer and Schuhfried recorded an impedance signal similar to that shown in Figure 1. Classically, these traces have been represented mirrored horizontally, so that the Y axis represents blood volume instead of impedance. They observed that, following the QRS complex, the REG trace increases abruptly suggesting a sudden blood inflow into the cranial cavity. Subsequently, a

Figure 1. Example of a REG signal and an ECG signal obtained in our laboratory



slow descent in impedance brings the REG trace back to its original value, and it is a result of the constant venous outflow. The morphological differences between the REG signals of pathological and normal subjects suggested that REG trace could be caused by the CBF pulsatility.

Nevertheless, despite the medical necessity of a device to evaluate the CBF, REG ability to evaluate the CBF was strongly debated. Detractors argued that the relative low electrical conductivity of the skull causes most of the current injected from the scalp surface to flow through the scalp tissue without crossing the skull. Accordingly, most of the REG information would be of extracranial origin.

The interest in the technique, and the disagreement about the origin of the REG signal, led to extensive research on the topic. For example, a review of the literature classified under the heading of rheoencephalography between 1959 and 1964 reveals some 150 articles, and one book (Namon, Gollan, Shimojyo, Sano, Markovich, & Scheinberg, 1967). Jenkner (1962) published a monograph summarizing his findings on more than 400 animal experiments, and well over 6,000 clinical records, and compared the results with those obtained by other authors. Some other exhaustive literature reviews can also be found in Lechner and Martin (1967), Lifshitz (1970), and Hadjiev (1972).

On the contrary, Perez-Borja and Meyer (1964) and Masucci, Seipel, and Kurtzke (1970) did not find significant differences in the REG signals recorded from normal subjects and from patients with cerebrovascular diseases and neurological deficits; Laitinen (1968) compared REG signals with intracerebral impedance traces, and found them to be substantially different; Masucci et al. (1970) applied the quantitative method proposed by Seipel (1967) to pathological and normal subjects, and found no clinical value of that analytical method; and, finally, Waltz and Ray (1967) analyzed the REG traces recorded from 96 subjects, some of them with partial or complete obstruction of an internal carotid artery or recent ischemic cerebral infarcts, and considered REG inadequate for the clinical evaluation of cerebrovascular disorders.

The interest in REG and its research flagged in the ‘70s at the same time as transcranial Doppler ultrasound appeared. Nevertheless, despite the lack of agreement on the REG signal origin, some other research groups have still contributed to REG knowledge during the following decades. For instance, Jacquy reported several findings about the use of REG in the study of regional CBF (Jacquy, Dekoninck, Piraux, Calay, Bacq, Levy, & Noel, 1974; Jacquy, Piraux, Jocquet, Lhoas, & Noel, 1977a; Jacquy, Piraux, Jocquet, Lhoas, & Noel, 1977b; Jacquy & Squelart , 1984), and Shender and Dubin (1994) used REG in the study of the CSF displacement on acceleration stress in military aircrafts. Meanwhile, Hatsell (1991) showed mathematically that regional impedance changes in the cortex are hardly observed by REG, whereas Basano et al. (2001) did not find significant differences in the REG signals recorded from brain-dead patients, and those obtained from living patients. In summary, research on REG ability to reflect the intracranial blood flow has coexisted with the use of REG in clinical studies.

REG MODELING

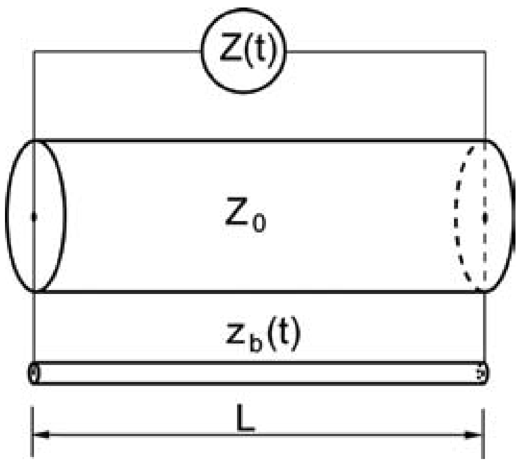
In engineering, models have been extensively used to study theoretical aspects involved in physical processes. In general, the more complex the model is, the closer to reality the results are. On the contrary, a simple model allows the study of each variable separately and independently from all the others. In this sense, only a few theoretical studies about REG ability to inspect the intracranial blood flow were published during the early stages in REG research.

To determine the changes in impedance caused by the blood inflow into an extremity, Nyboer modelled limbs as a cylinder with a constant impedance Z_0 connected in parallel to a second variable-diameter cylinder in such a way that impedance $z_b(t)$ is time-dependent. This second cylinder represents the net blood that flows in the extremity’s tissue along time (Figure 2). This simple model provides a relation between the blood volume increment (ΔV), and the impedance increment (ΔZ) as:

$$\Delta V = -\rho_b \left(\frac{L}{Z_0}\right)^2 \Delta Z$$

where ρ_b is the blood resistivity, L is the interelectrode distance, and Z_0 is the basal impedance of the limb (Geddes & Baker, 1989; Nyboer, 1970). Unfortunately, this simple model is not appropriate for REG analysis, since a portion of the current injected from the scalp surface crosses several tissue layers (e.g., scalp, skull,

Figure 2. Schematic diagram of Nyboer's model used in plethysmographical studies; the upper cylinder represents the impedance of the extremity, and lower cylinder corresponds to the time dependent impedance of the net blood flowing into the extremity



cerebrospinal fluid (CSF), and brain), whereas another portion flows through the scalp tissue only.

On the other hand, specific models for REG analysis have also been developed. For instance, Namon et al. (1967) and Seipel (1967) proposed the use of a multilayer sphere for head modeling, including regions corresponding to scalp, skull, cerebrospinal fluid, and brain. This simple head model is still being used in many applications, because it allows analyzing independently the influence of each geometrical parameter or electrode localization on the results, and even deriving the analytical expression of the solution (e.g., Reyhani & Ludwig, 2006). In addition, Lifshitz (1970) compared the distribution of the equipotential lines resulting from the use of different current injection electrode diameters by means of a two-dimensional head model consisting of three concentric circles of resistive Teledeltos paper. Today, powerful computer systems have allowed the development of multilayer spherical models of the head, providing the corresponding numerical and analytical solutions (Perez, Guijarro, & Barcia, 1999; Perez, Guijarro, & Barcia, 2000; Perez, Guijarro, & Barcia, 2004).

For a sound analysis of REG's theoretical principles, besides the geometrical factors of the head model, it is necessary to take into account how REG impedance is measured from the scalp surface. In practice, a pair of electrodes is attached to the scalp, and a small current is injected through them. Amplitudes of 1 or 2 mA and frequencies of 20 kHz to 100 kHz are usual values for the injected current. The diameter of the current injection electrodes is also of great importance, since the amount of current crossing the cranium can be significantly increased by using large electrodes (Namon et al., 1967). This current flow generates an electric field whose potential can be measured from the scalp surface. If the electric potential is measured between the current injection electrodes, the REG is said to be bipolar or REG I. This electrode arrangement is very sensitive to movements in the electrode leads and to noise, since most of the measured impedance is caused at the electrode-skin interface. Alternatively, a second pair of electrodes can be used to measure the electric potential difference created by the current injection, known as tetrapolar REG or REG II. Some other interesting REG electrode arrangements and techniques can be found in Namon et al. (1967) and Lifshitz (1967).

Based on geometrical head models, Perez et al. (2000) analyzed the ability of REG I and REG II to reflect the intracranial blood flow. They modelled the head as four concentric spheres representing, from the innermost to the outermost, brain, CSF, skull and scalp (Figure 3). The current injection electrodes were modelled as two high-conductivity cylinders attached to the scalp in such a way that the scalp volume intersected the electrodes.

This theoretical study only analyzed REG signal recoded from electrode arrangements totally symmetrical to the horizontal plane and for some electrode positions, in which the angle between the radial vector of the current injection electrodes (α) and the angle between the radial vector of the impedance reading points (β) (pickup electrodes) changed (Figure 3a). In this work, the head model was meshed and solved by the finite element method (Figure 3b). Observe that the electrode configuration studied there is a REG I when the current injection electrodes and pickup electrodes match (α=β); otherwise, it is a REG II.

The main results of the model used in Perez et al. (2000) are represented in Figure 4; the unitary contribution of the intracranial blood circulation to REG (M_1 ratio) is represented as a function of the electrode arrangement. Given a REG measurement, the M_1 ratio represents the amount of impedance per unit that can be attributed to the CBF. According to this study, most of

Figure 3. Geometrical head model used by Perez et al. (2000) to study REG ability to reflect the cerebral blood flow; (a) schematic diagram of the model; (b) meshed model

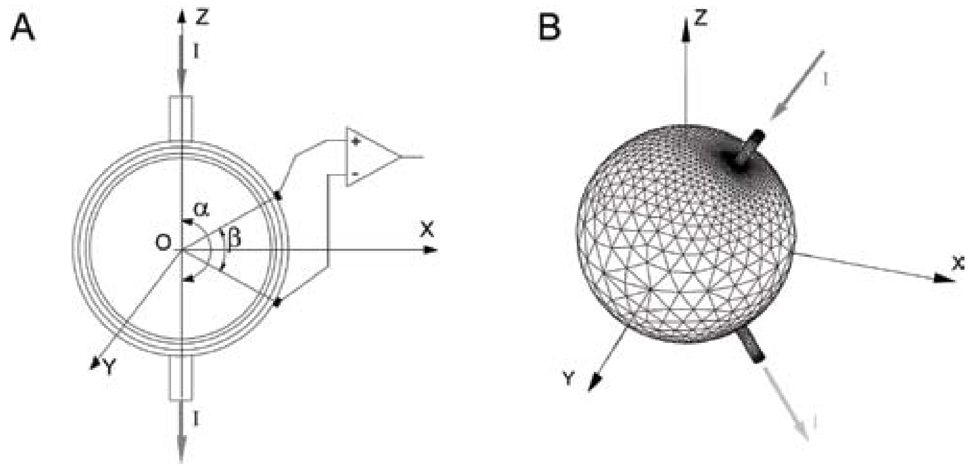
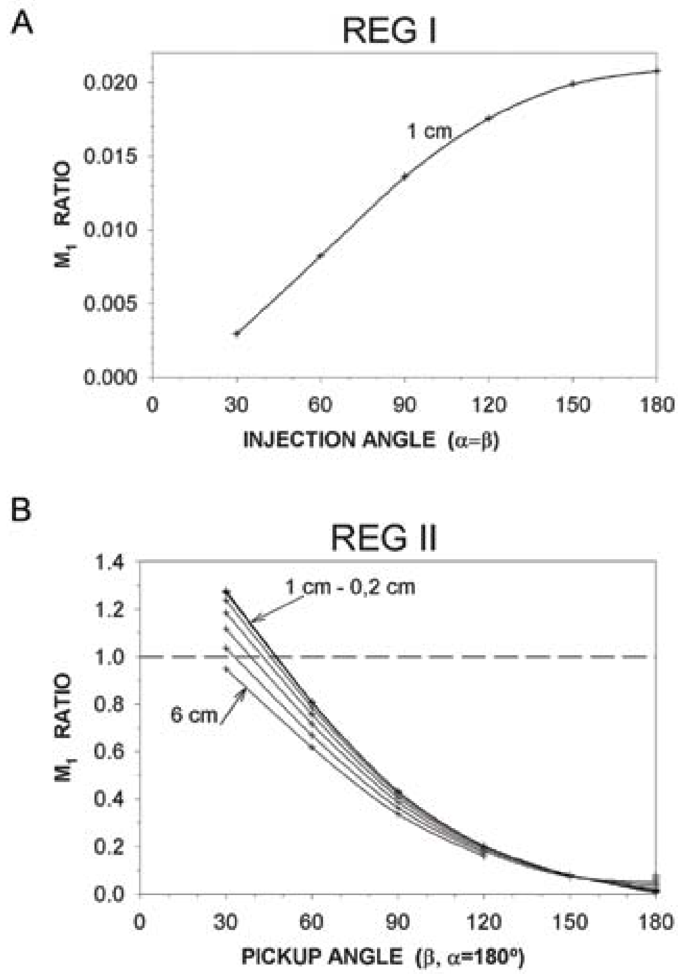


Figure 4. Relative contribution of intracranial circulation to REG (M_1); (A) M_1 ratio for REG I as a function of the injection angle α; (B) M_1 for REG II as a function of the pickup angle β



the REG I trace is caused by the scalp blood flow, since scarcely 2% of intracranial information may appear in the REG trace when the current injection electrodes are placed at diametrically opposite sites (α=180º).

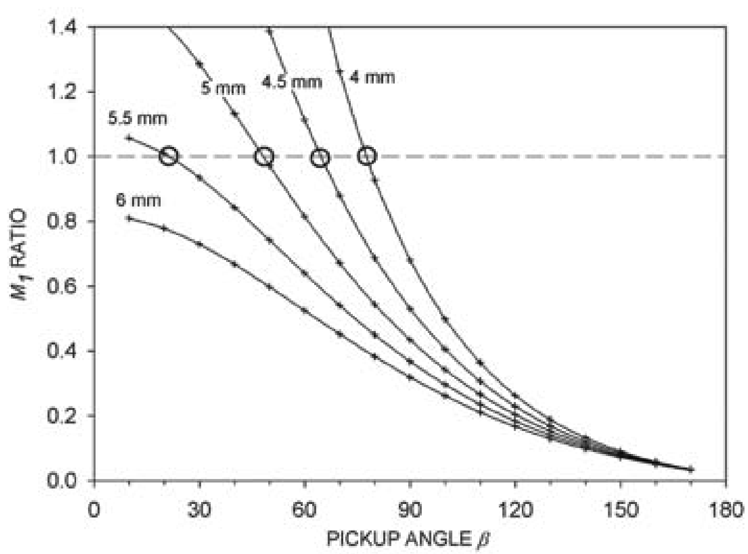
However, the analysis of the M_1 ratio in REG II reveals that, theoretically, there are some electrode arrangements from which a REG II signal free of extracranial contribution can be recorded (M_1=1). Obviously, the model is a simple approximation to reality, and hence, one should not expect to obtain a pure intracranial REG simply by placing the current injection electrodes at sites similar to those used in the model. However, these results suggest that the finding of an electrode arrangement to register a REG II free of extracranial contribution is physically possible, despite the fact that some of the injected current flowed only through the scalp tissue. Moreover, REG II signal results also show that the diameter of the current injection electrodes has only a small influence on M_1, and, in any case, it only slightly modifies the optimal electrode position. As a consequence, the use of large current injection electrodes in REG II experimental studies seems unnecessary.

In practice, a very high level of expertise is required to find an optimal electrode arrangement for a given subject to record a REG II free of extracranial contribution, only by inspecting the morphology of the trace. This is because the morphology of the intracranial REG component is unknown, and can only be observed by the detection of changes in the slope of the extracranial REG component, whose morphology is easily recognized.

In this work, the head model was solved by the finite element method, and although this method allows analyzing complex geometrical structures, any change in the geometrical parameters means the remeshing and resolution of the model again. Recently, this work has been extended to the study of the influence of the scalp thickness on REG II (Perez et al., 2004), and in this study, the analytical expression of the potential electric field created by the current injection for α=180º was derived.

The main results reported in Perez et al. (2004) are shown in Figure 5, in which the M_1 ratio is plotted as a function of the pickup angle β for different scalp thicknesses. Unexpectedly, the optimal electrode arrangement to obtain a REG II exclusively caused by intracranial circulation (M_1=1) depends strongly on the scalp thickness. Furthermore, a unitary M_1 ratio is not reached for scalp thickness values greater than 6 mm. This means that the search of an optimal electrode arrangement could be successful in some subjects, and fruitless in other subjects. Or, in other words, given an electrode arrangement, the REG signal recorded can either contain pure intracranial information, or be a course mixture of both extra and intracranial information, depending on the subject's physical constitution.

Figure 5. Relative contribution of intracranial circulation to REG (M_1) as a function of pickup angle (β) and scalp thickness, for a 180 º current injection angle (α)



The quantitative results reported in these works (Perez et al., 1999; Perez et al., 2000; Perez et al., 2004) cannot be directly compared to other experimental results, since they have been derived from simple geometrical models. However, they allow us to draw some interesting conclusions that agree well with earlier experimental works: (i) REG I mainly reflects extracranial information; (ii) the REG II signal is a mixture of information from intra and extracranial blood flow in unknown proportions; and (iii) there is no universal electrode arrangement to register a REG II free of extracranial contribution suitable for all individuals.

A close review of the literature shows comparative studies based on different electrode arrangements, and even different REG techniques. Furthermore, some reviews do not mention the REG technique used by other authors in their experiments. Both the comparison of results derived from different REG techniques, along with the unforeseeable amount of intracranial information contained in a REG signal, could be the cause of disagreement about ability of REG to reflect the intracranial blood flow.

FUTURE TRENDS

REG has not been included in the clinical practice yet, because the physical principles of this technique were not wholly understood. This caused the aforementioned disagreement that led to discarding further research on REG. Now, once the REG basis has been determined, the search of a method to eliminate the extracranial artifact from the REG signal can be addressed. For instance, Perez, Guijarro, and Sancho (2005) have recently observed that the extracranial component of the REG signal can be considered to be morphologically constant along the scalp surface. This means that the pattern of the extracranial artifact can be easily recognized by recording, for example, a REG I, and this information can be removed from a REG II signal to obtain a pure intracranial impedance signal (Perez, Guijarro, Sancho, & Navarré, 2006).

CONCLUSION

Despite the technological advances in tomographic imaging, the idea of a noninvasive, portable, and low-cost device to monitor the cerebrovascular system still seems attractive. REG signal was extensively studied and clinically used for this purpose some years after its proposal, but it was finally neglected because its theoretical principles were not firmly established. However, a deeper understanding of this technique would allow the recovery of REG to the clinical practice.

N

ACKNOWLEDGMENT

This work was supported by grant PI04/0303 from the Instituto de Salud Carlos III (Fondo de Investigación Sanitaria) in the framework of the “Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica (I+D+I).”

REFERENCES

Basano, L., Ottonello, P., Nobili, F., Vitali, P., Pallavicini, F.B., Ricca, B., Prastaro, T., Robert, A., & Rodríguez, G. (2001). Pulsatile electrical impedance response from cerebrally dead adult patients is not a reliable tool for detecting cerebral perfusion changes. *Physiological Measurement*, *22*(2), 341–349.

Geddes, L.A., & Baker, L.E. (1989). *Principles of applied biomedical instrumentation*. New York: Wiley.

Hadjiev, D. (1972). Impedance methods for investigation of cerebral circulation. *Progress in Brain Research*, *35*, 25–85.

Hatsell, C.P. (1991). A quasi-power theorem for bulk conductors: Comments on rheoencephalography. *IEEE Transactions on Biomedical Engineering*, 38(7), 665–669.

Jacquy, J., Dekoninck, W.J., Piraux, A., Calay, R., Bacq, J., Levy, D., & Noel, G. (1974). Cerebral blood flow and quantitative rheoencephalography. *Electroencephalography and Clinical Neurophysiology*, *37*(5), 507–511.

Jacquy, J., Piraux, A., Jocquet, P., Lhoas, J.P., & Noel, G. (1977a). Cerebral blood flow in the adult blind: A rheoencephalographic study of cerebral blood flow changes during Braille reading. *Electroencephalography and Clinical Neurophysiology*, *43*(3), 325–329.

Jacquy, J., Piraux, A., Jocquet, P., Lhoas, J.P., & Noel, G. (1977b). Regional cerebral blood flow during voluntary movements of the hand. A rheoencephalography study. *Neuropsychobiology, 3*(4), 240–249.

Jacquy, J., & Squelart, P. (1984). Right hemispheric activation on chronognosia: A preliminary study by rheoencephalography. *Neuropsychobiology*, *11*(1), 3–6.

Jenkner, F.L. (1962). *Rheoencephalography. A method of the continuous registration of cerebrovascular changes*. Springfield: Charles C Thomas.

Laitinen, L.V. (1968). A comparative study on pulsatile intracerebral impedance and rheoencephalography. *Electroencephalography and Clinical Neurophysiology*, *25*(3), 197–202.

Lechner, H., & Martin, F. (1967). Development of the rheoencephalographic investigation methods. *Electroencephalography and Clinical Neurophysiology. Supplement, 25*, 119–127.

Lifshitz, K. (1970). Electrical impedance cephalography (rheoencephalography). In M. Clynes, & J.H. Milsum (Eds.), *Biomedical Engineering Systems* (pp. 21–64). New York: McGraw-Hill.

Malmivuo, J., & Plonsey, R. (1995). *Bioelectromagnetism*. New York: Oxford University Press.

Masucci, E.F., Seipel, J.H., & Kurtzke, J.F. (1970). Clinical evaluation of "quantitative" rheoencephalography. *Neurology, 20*(7), 642–648.

Namon, R., Gollan, F., Shimojyo, S., Sano, R.M., Markovich, S.E., & Scheinberg, P. (1967). Basic studies in rheoencephalography. *Neurology*, *17*(3), 239–252.

Nyboer, J. (1970) *Electrical impedance plethysmography*. Springfield: Charles C. Thomas.

Perez, J.J., Guijarro, E., & Barcia, J.A. (1999). Quantification of sensitivity to conductivity changes in bipolar rheoencephalography. In *Proceedings of the First Joint BMES/EMBS Conference* (Vol. 1, pp. 455).

Perez, J.J., Guijarro, E., & Barcia, J.A. (2000). Quantification of intracranial contribution to rheoencephalography by a numerical model of the head. *Clinical Neurophysiology*, *111*(7), 1306–1314.

Perez, J.J., Guijarro, E., & Barcia, J.A. (2004). Influence of the scalp thickness on the intracranial contribution to rheoencephalography. *Physics in Medicine and Biology*, *49*(18), 4383–4394.

Perez, J.J., Guijarro, E., & Sancho, J. (2005). Spatio-temporal pattern of the extracranial component of the rheoencephalographic signal. *Physiological Measurements*, *26*(6), 925–938.

Perez, J.J., Guijarro, E., Sancho, J., & Navarré, A. (2006). Extraction of the intracranial component from the rheoencephalographic signal: A new approach. In *Proceedings of the 28th IEEE – EMBS Annual International Conference* (Vol. 1, pp. 6064–6067).

Perez-Borja, C., & Meyer, J.S. (1964). A critical evaluation of rheoencephalography in control subjects and in proven cases of cerebrovascular disease. *Journal of Neurology, Neurosurgery, and Psychiatry*, *27*, 66–72.

Polzer, K., & Schuhfried, F. (1950). Rheographische untersuchungen am schädel. *Wiener Zeitschrift fur Nervenheilkunde und deren Grenzgebiete*, *3*, 295–298.

Reyhani, S.M., & Ludwig, S.A. (2006) An implanted spherical head model exposed to electromagnetic fields at a mobile communication frequency. *IEEE Transactions on Bio-Medical Engineering*, *53*(10), 2092–2101.

Shender, B.S., & Dubin, S.E. (1994). Contribution of cerebrospinal fluid to rheoencephalographic waveforms during hypoxic and +Gz stress. *Aviation, Space, and Environmental Medicine*, *65*(6), 510–517.

Seipel, J.H. (1967). The biophysical basis and clinical applications of rheoencephalography. *Neurology*, *17*(5), 443–451.

Waltz, A.G., & Ray, C.D. (1967). Inadequacy of "rheoencephalography." A clinical study of impedance cephalography for evaluation of cerebrovascular disorders. *Archives of Neurology*, *16*(1), 94–102.

KEY TERMS

Bipolar Rheoencephalogram (REG I): Pulsatile impedance changes of the head, related to the cardiac cycle that is measured by injecting a small current between two electrodes attached to the scalp surface, and reading the electric potential difference generated between them.

Electrical Conductivity: Measure of the ability of a material or substance to conduct an electric current.

Electrical Impedance: Measure of the opposition to the flow of a sinusoidal alternating electric current.

Engineering Model: Simplification of the reality to address a complex problem by simple rules that can be described by means of equations.

Finite Element Method: Mathematical technique to solve complex problems in which a physical domain is spatially discretized (meshed) and solved by numerical methods.

Rheoencephalography: Electromedical technique used to assess the cerebral blood flow (CBF) by noninvasive electrical impedance methods using electrodes attached to the scalp surface.

Tetrapolar Rheoencephalogram (REG II): Pulsatile impedance changes of the head, related to the cardiac cycle that is measured by injecting a small current between two electrodes attached to the scalp surface, and reading the electric potential difference between another pair of electrodes.

A Non-Invasive Approach to the Bionic Eye

N. Grossman
Imperial College London, UK

H. Yang
Imperial College London, UK

E. Drakakis
Imperial College London, UK

K. Nikolic
Imperial College London, UK

P. Degenaar
Imperial College London, UK

C. Toumazou
Imperial College London, UK

INTRODUCTION

According to the World Health Organisation definition for blindness, that is, visual acuity bellow 3/60 for the best eye on the Snellen scale, there are thought to be 38 million blind people worldwide (Delbeke et al., 2004). This figure is expected to double over the next 25 years due to combination of an increasing population and aging worldwide. There are additionally 110 million people who have severely impaired vision and are high risk of becoming blind. The most common causes of blindness are: cataract, trachoma, glaucoma, diabetic retinopathy, age related macular degeneration (AMD) and retinitis pigmentosa (RP). In the west countries, cataract and glaucoma make up only 11% of the total causes of blindness. In these regions AMD and RP are prevalent eye diseases. AMD increases dramatically with age, so that (with about 2million cases in the USA) it is the leading cause of blindness among Americans of European descent (Friedman et al., 2004).

The AMD and RP result in the loss of photosensitivity primarily due to destruction of the rod and cone photoreceptors. Medical intervention to date has been disappointing. There is no known mechanism by which the eye can self-repair. Anti-angiogenesis drugs can significantly slow down the progression of wet type AMD, but in most cases there is very little treatment. Even more significantly, none of the drugs are capable of restoring lost vision. The idea of using stem cells in therapies is still complex and may be many decades away from potential treatment. Prosthetic implants are therefore the only method at present by which we can offer a return of some of the lost vision. Here we present a special type of vision restoration based on the optical stimulation of retinal ganglion cells (RGCs), which remain operational.

BACKGROUND

Most of current concepts for a visual prosthesis are based on neuronal electrical stimulation. Several different locations along the visual pathways have been proposed and examined so far for visual prosthesis:

- Retinal implant (which is either epiretinal or subretinal)
- Optic nerve implant and
- Visual cortex implant

A nice historical perspective regarding electrical visual prosthesis can be found in Lovell et al. (2007). There is a number of very good reviews of the artificial vision and visual prosthesis, such as Dowling (2005), Loewenstein, Montezuma, and Rizzo (2004), Margalit et al. (2002), Zrenner (2002), and so forth. The most favoured approach is the intraocular electrode array, since the retinal implant seems to be the most convenient for surgical intervention and it is easier to retinotopically map individual pixels on the light sensing array with points on the retina.

In most forms of retinal dystrophy, the photoreceptors are lost, but the RGCs, the output neurones of the retinal network, survive and project to the retinorecipient areas of the visual cortex. There have been numerous attempts to stimulate these cells as a prosthetic strategy. The implants to date have come in two forms: subretinal and epiretinal. Subretinal implants are placed underneath the retinal layers and usually consist of a microphotodiode array which attempts to stimulate the remaining signal processing layers in the retina (U.S. Patent No. 5,024,233, 1991). The epiretinal chips are positioned on the surface of the retina, and try to stimulate the retinal ganglion cell layer (U.S. Patent No. 5,935,155, 1999). In this case, additional image processing is required to replicate that of the bypassed retinal layers.

The retinal prosthetics that have been implanted to date have several problems (Dowling, 2005). Both types of implants have to be implanted inside the eyeball (and subretinal structures need to be placed under the retina, requiring more delicate surgery) and the stimulating electrodes have to be in good physical contact with the cells which they stimulate. The invasive electrodes can cause further degradation of remaining functional tissue. The curved nature of the eye makes it very difficult to cover any significant portion of the retina, though there have been attempts at more flexible substrates. Furthermore, power consumption issues are highly significant. Introduction of power cables is difficult, and power transmission through an RF link, for example, does not always provide sufficient power to power a significant quantity of stimulation points. Additionally, the position of the electrodes is fixed once they are inserted, and precise positioning on the micrometer scale is hard to control.

To date, epiretinal implants have used small numbers of electrodes and thus the patient has seen light flashes known as phosphenes. It has been possible for patients to interpret some basic shapes from these phosphenes (Humayun et al., 2003). However, if the stimulation matrix is to be scaled to allow detailed analysis of the visual scene, then the image processing component will become increasingly important.

There has been substantial progress toward an electronic retinal prosthesis, but fully functional, long-lasting devices are presently limited largely by power consumption issues.

PHOTONIC RETINAL PROSTHESIS

Optical Stimulation of Neuron Cell

Recent advances in biochemistry have generated a novel neuronal stimulation technology that is based on light. Using light to stimulate neuronal signals has many advantages. First, in contrary to electrode based stimulation, light does not require mechanical contact and can enable noninvasive remote control on the cell activities. Second, since light beams can be easily scanned, the stimulating point is flexible and does not require a prior decision on the stimulation location. Third, using conventional optics light can be focused to very high spatial resolution.

The feasibility of using light instead of electricity to trigger neuronal signals was first demonstrated in 1971 by Richard Fork who used high power laser to stimulate action potentials in the abdominal ganglion of the marine mollusk Aplysia California (Fork, 1971). Since then scientists have been exploiting developments in Nanotechnology and Genomics to realize various photostimulation tools that would revolutionize biology and neuroscience. The impartment of light sensitivity, also known as photosensitization, is achieved by either:

1. Photolysis of blocking moiety attached to neurotransmitters
2. Expression of natural photosensitive proteins and ion channels
3. Attachment of a photoisomerizable switch next to the active site of ion channel or membrane receptor.

Photolysis of Caged Neurotransmitter

The photolysis of caged neurotransmitter approach is based on disabling the neurotransmitter activity by attaching a photoremoveable blocking compound. This idea was first introduced in 1978 by Kaplan, Forbush, and Hoffman (1978), but the utility of the caged compounds tool became widely apparent to the neuroscience world just in 1986 when an NB-caged-acetylcholine agonist, carbamoylcholine, was used to study the kinetics of nicotinic acetylcholine receptors (Walker, McCray, & Hess, 1986). Since then photolysis of caged molecules has gained a wide use as a remote control means of stimulating cellular responses. Since all CNS neuron cells express glutamate receptors, caged glutamate is the most widely used caged neurotransmitter (Callaway & Katz, 1993).

The uncaging process has very good temporal resolution. A temporal spike jitter, which is the time between stimulus and the first single spike, was reported to be reliably evoked with 1-3 ms (Shoham, O'Connor, Sarkisov, & Wang, 2005). The spatial resolution is more limited since it is influenced by the diffusion of the uncaged molecules. However, recently were reported spatial patterns of up to 5 μm (Shoham et al., 2005).The remarkable resolution of the caged neurotransmitters technique enables the studies of single dendrite interactions and mapping of neuronal connectivity (Yoshimura, 2005). The downsides of the technique are:

- The stimulation of unintended sites and reduction of the repeatability of the process due to diffusion of the neurotransmitter,
- It is difficult to target single subpopulation in heterogeneous tissue due to the wide sensitivity to glutamate,
- The process requires continual supply of the caged molecules,
- The UV excitation reduces the biocompatibility of the approach and hinders stimulating in volumetric tissue.

Expression of Natural Photoreceptive Protein

Expression of natural photoreceptive proteins is probably the most straight forward approach to render light sensitivity in cells. Here we also include direct addition of photosensitive proteins, such as Photosystem-I (Greenbaum et al., 2002). Natural photosensitive proteins have the advantage of being optimized by evolution over millions of years and they have good quantum efficiency. Most of the photosensitive proteins in the eukaryotic cells are in the same family as rhodopsin, the protein found in the retina rods and cones cells. Rhodopsin and its relatives all have an apoprotein (opsin) and covalently linked a photoisomerizable chromophore (a light sensitive molecule such as 11-cis retinal). The *cis-trans* photoisomerization of the chromophore triggers a conformational change in the protein that signals downstream cellular events through G protein-coupled cascade (Kramer, Chambers, & Trauner, 2005). The feasibility of using a natural (invertebrate) light sensitive system was first demonstrated in 2002 (Zemelman, Lee, Ng, & Miesenbock, 2002). They showed that non sensitive cells can be converted to respond to light by expressing only three proteins of the phototransduction cascade: rhodopsin, arrestin-2, and α subunit of G-protein. However, this approach had poor control (jitter length in the order of seconds/minutes) and bad reproducibility. Since then, two additional photosensitive proteins have been developed to render neural cells light sensitive.

The first one was melanopsin which is a photopigment that was discovered in small subset (~2%) of intrinsically light-sensitive retinal ganglion cells that are not involved in the image formation process, but instead regulate the circadian system and the pupil constriction (Foster, 2002). The phototransduction in the melanopsin is based on an invertebrate type G-protein cascade. Melyan, Tarttelin, Bellingham, Lucas, and Hankins (2005) showed that when melanopsin was expressed in mammalian cell lines it enabled light-evoked depolarization responses. The spike jitter of the melanopsin can be seconds to minutes long (Qiu et al., 2005). Melanopsin has a sensitivity maximum at 480nm, and the threshold light intensity required for producing responses is near 1012 photons cm-2s-1. For comparison, the thresholds for the rod and cone photoreceptors are about 106 and 1010 photons cm-2s-1. The advantage of the melanopsin is that it is bistable due to its intrinsic capability to reisomerise the retinal molecule back to its cis state after the photoisomerization (Melyan et al., 2005).

The second photosensitive protein is the algal Channelrhodopsin-2 (ChR2). When expressed in animal cells ChR2 forms small-conductance, nonselective cation channels that can be activated by exposure to blue light (Boyden, Zhang, Bamberg, Nagel, & Deisseroth, 2005). Its photosensitivity is based on photoisomerization of all-trans retinal to a cis conformation. Like melanopsin, the retinal in ChR2 can be reisomerised to back to its *all-trans* state alleviating the need for external enzyme (Bi et al., 2006). Expression of ChR2 in the surviving inner retinal neurons of mice with photoreceptors degeneration restored the ability of the retina to detect light signals and transmit it to the visual cortex (Bi, 2006). ChR2 has a sensitivity peak in blue 470nm wavelength. Its activation threshold was reported to be between 1015 photons cm-2s-1 (Bi et al., 2006) and 1018 photon cm-2s-1 (Zhang et al., 2006). Contrary to melanopsin and rhodopsin, the phototransduction in ChR2 does not involve any amplifying signalling cascades, since the ChR2 is sensor and ion channel combined. Hence, cells that express ChR2 respond very fast. The channel is opened within just 50 µs from illumination and first spikes are expected after 1-3 ms. Boyden et al. (2005) used this unique kinetics to demonstrate for the first time a reliable millisecond timescale triggering of single action potentials (Boyden et al., 2005). Further investigations of the kinetics of ChR2 were published in Nikolic, Degenaar, and Toumazou (2006).

Synthetic Photoswitch

Photoswitching is also possible by coupling a synthetic photoisomerizable nano compound to an ion channel or a receptor protein. The azobenzene molecule is an example for such a compound. It shows a *trans*-to-*cis* photoizomerization upon absorption of a photon of the energy 3.4eV (λ = 365nm) and *cis*-to-*trans* at 1.95eV (λ = 420nm). This property has been recently used to form light controlled shaker channel (Banghart, Borges, Isacoff, Trauner, & Kramer, 2004). The uniqueness of the photswitch approach is its versatility since the photoswitch is not limited to a specific ion channel. At present, several limitations hinder the success of the technique: higher energy *cis* state is thermally isomerised to *trans* state disabling the bistable switch, the switching involve UV light, the photoswitch hyperpolarizes the cells and hinder action potential and the current method involves considerable genetic engineering.

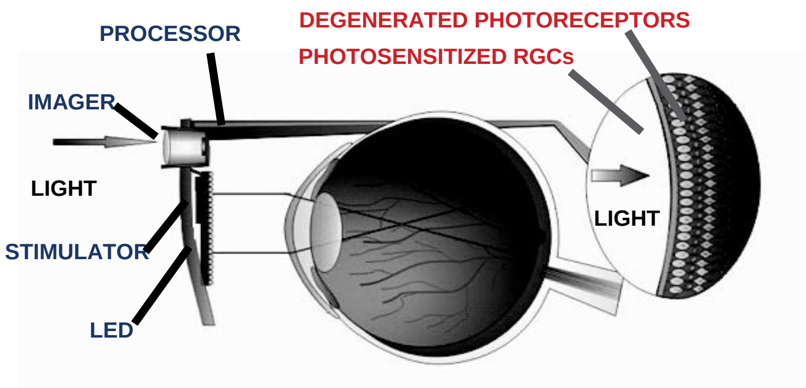
Description of the Prosthesis

The photonic retinal prosthesis, shown in Figure 1, is based on photosensitization of the residual retinal ganglion cells. The device comprises an image capturing component, light producing component, and control electronics to process the captured image and control stimulating waveforms of the light emission system (Degenaar et al., 2006). The light sources may be LEDs or lasers, for example, vertical cavity lasers. Each emitter will require individual control thus dedicated optoelectronic systems are required (Yan, Drakakis, Toumazou, Nikolic, & Degenaar, 2006), rather than typical raster scanned displays. The choice of the array will therefore depend on integration with processing electronics and power consumption.

The image processing is similar to that required for epiretinal devices, as the bypassed retinal processing needs to be replicated. Typically Laplacian filtering of the image is sufficient to replicate spatial processing (Banks, Degenaar, & Tournazou, 2005), but the full spatiotemporal processing will be more complex. Once the image is converted into the required retinomorphic format, individual points on the image matrix will need to be convolved with a stimulating transform to acquire the required stimulating waveform matrix. This will then be used to control the LED components which will emit the required optical waveforms to stimulate the RGC's and thereby the visual cortex.

Electrophysiological recordings of ganglion cells reveal diverse spiking patterns with maximal frequency up to several tens of kHz (Rullen, 2001). Additionally, ChR2, which is presently the most advanced candidate for photosensitization, will also have a variance in expression. Thus, the optical stimulation system will need to be tunable to cope with this variance. There is still a variance in the published literature about the required light intensities required for ChR2 stimulation. These vary between 0.1 and 100 mW/cm2. Given that

Figure 1. Outline of photonic retinal prosthesis device



functional vision restoration will require about 1cm2 of retinal stimulation, then assuming 5% LED quantum efficiency and 10% coupling efficiency, and negligible relative power consumption of the control circuitry, the final device will consume between 20mW and 20W. The former will last about 150 hours on a typical mobile phone battery, the latter about 10 minutes. It is thus crucial to maximize the photosensitivity of the RGCs and the efficiency of the light emission system.

FUTURE TRENDS

Optoelectronic vision restoration is presently in its infancy. As highlighted previously, there is still research to be done on maximizing the photosensitivity of the Retinal Ganglion cells. Additionally, the optical stimulation systems and control electronics will need development. The present potential directions are with flip-chip GaN based arrays on CMOS chips and organic/CMOS hybrid emissive systems that are used in present OLED displays. The latter is much more efficient, but may not be able to obtain the required brightness.

CONCLUSION

As this research progresses from the Petri dish to the animal models to humans, much will be learned not only about vision restoration but about the nature of communication with the visual cortex, and the adaptability of the visual cortex to new forms of image interpretation. Thus, the clinical implementation of this technology is around a decade from becoming commercial reality, but nevertheless provides exciting hope for the future.

REFERENCES

Banghart, M., Borges, K., Isacoff, E., Trauner, D., & Kramer, R. (2004). Light-activated ion channels for remote control neuronal firing. *Nature Neuroscience, 7*(12), 1381-1385.

Banks, D.J., Degenaar, P., & Tournazou, C. (2005). Distributed current-mode image processing filters. *Electron Lett., 41*, 1201-1202.

Bi, A., Cui, J., Ma, Y.P., Olshevskaya, E., Pu, M., Dizhoor, A.M., et al. (2006). Ectopic expression of a microbial-type rhodopsin restores visual responses in mice with photoreceptor degeneration. *Neuron, 50*, 23-33.

Boyden, E., Zhang, F., Bamberg, E., Nagel, G., & Deisseroth, K. (2005). Millisecond-timescale, genetically targeted optical control of neural activity. *Nat. Neurosci., 8*, 1263-1268.

Callaway, E.M., & Katz L.C. (1993). Photostimulation using caged glutamate reveals functional circuitry in living brain slices. In Proceedings of the Natl. Acad. Sci. (pp. 7661-7765).

Chow, A.Y. (1991). U.S. Patent No. 5,024,223. Washington, DC: U.S. Patent and Trademark Office. Artificial retina device.

Degenaar, P., Hankins, M.W., Drakakis, E., Toumazou, C., Kennard, C., Nikolic, K.. et al. (2006). *Non-invasive retinal prosthetic*. Patent application GB 0612242.8, filed 21/6/2006.

Delbeke, J. et al. (2004). Vision rehabilitation in the case of blindness. *Expert Rev. Medical Devices, 1*(1), 139-53.

Dowling, J. (2005). Artificial human vision. *Expert Review of Medical Devices, 2*(1), 73-85.

Fork, R.L. (1971). Laser stimulation of nerve cells in Aplysia. *Science, 171*, 907-908.

Foster, R.S. (2005). Bright blue times. *Nature, 433*, 698-699.

Friedman, D.S. et. al. (2004). The eye diseases prevalence research group. Prevalence of age-related macular degeneration in the United States. *Archives of Ophthalmology, 122*, 564-572.

Greenbaum, E., Humayun, M., Kuritz, T., Lee, J.W., Sanders, C.A., Bruce, B., et al. (2002). Nanoscale photosynthesis, the photophysics of neural cells and artificial sight. In *Proceedings of the IEEE-EMBS Special Topic Conference on Molecular, Cellular and Tissue Engineering* (pp. 83-85).

Humayun, M.S., de Juan, E., & Greenberg, R.J. (1999). U.S Patent No. 5,935,155. Washington, DC: U.S. Patent and Trademark Office. Visual prosthesis and method of using same.

Humayun, M.S., Weiland, J.D., Fujii, G.Y., Greenberg, R., Williamson, R., Little, J. et al. (2003). Visual perception in a blind subject with a chronic microelectronic retinal prosthesis. *Vision Research, 43*(24), 2573-2581.

Kaplan, J.H., Forbush, B., III, & Hoffman, J.F. (1978). Rapid photolytic release of adenosine 5'-triphosphate from protected analogue: Utilization by the Na:K pump of human red blood cell ghosts. *Biochemistry, 17,* 1929-35.

Kramer, R.H., Chambers, J.J., & Trauner D. (2005). Photochemical tools for remote control of ion channels in excitable cells. *Nature Chemical Biology, 17,* 360-365.

Loewenstein, J.I., Montezuma, S.R., & Rizzo, J.F. (2004). Outer retinal degeneration - an electronic retinal prosthesis as a treatment strategy. *Archives of Ophthalmology, 122*(4), 587-596.

Lovell, N.H., Hallum, L.E., Chen, S.C., Socrates Dokos, P., Byrnes-Preston, R.G., Poole-Warren, L., et al. (2007). Advances in retinal neuroprosthetics. In Metin Akay (Ed.), *Handbook of Neural Engineering* (pp. 337-356). IEEE Press & Wiley-Interscience.

Margalit, E., Maia, M., Weiland, J.D., Greenberg, R.J., Fujii, G.Y., Torres, G., et al. (2002). Retinal prosthesis for the blind. *Survey of Ophthalmology, 47*(4), 335-356.

Melyan, Z., Tarttelin, E.E., Bellingham, J., Lucas, R.J, & Hankins, M.W. (2005). Addition of human melanopsin renders mammalian cells photoresponsive. *Nature, 433*, 741-745.

Nikolic, K., Degenaar, P., & Toumazou, C. (2006). Modelling and engineering aspects of ChannelRhodopsin2 system for neural photostimulation. In *Engineering in Medicine and Biology Society, 2006, 28th Annual International Conference of the IEEE* (pp. 1626-1629).

Qiu, X. et al. (2005) Induction of photosensitivity by heterologous expression of melanopsin. *Nature, 433,* 745-749.

Rullen, V.R., (2001). Rate coding versus temporal order coding: What the retinal ganglion cells tell the visual cortex. *Neural Comput***.,** *13*, 1255-1283.

Shoham, S., O'connor, D.H., Sarkisov, D.V., & Wang S.S. (2005). Rapid neurotransmitter uncaging in spatially defined patterns. *Nat. Methods, 2,* 837-843.

Walker, J.W., McCray, J.A., & Hess, G.P. (1986). Photolabile protecting groups for an acetylcholine receptor ligand. Synthesis and photochemistry of a new class of o-nitrobenzyl derivatives and their effects on receptor function. *Biochemistry, 25,* 1799-1805.

Yan, H., Drakakis, E., Toumazou, C., Nikolic, K., & Degenaar, P. (2006). An optoelectronic platform for retinal prosthesis. In *IEEE BioCAS, Biomedical Circuits and Systems Conference.*

Yoshimura, Y., Dantzker, J. L., & Callaway, E. M. (2005). Excitatory cortical neurons from fine-scale functional network. *Nature, 433,* 868-873.

Zemelman, B.V., Lee, G.A., Ng, M., & Miesenbock, G. (2002). Selective photostimulation of genetically chARGed neurons. *Neuron, 33*, 15-22.

Zhang, F., Wang, L.P., Boyden, E., & Deisseroth K. (2006). Channelrhodopsin-2 and optical control of excitable cells. *Nature Methods, 3*(10), 785-792.

Zrenner, E. (2002). Will retinal implants restore vision? *Science, 295,* 1022-1025.

KEY TERMS

ChannelRhodopsin/Melanopsin: proteins that can make cells light sensitive.

Implant: device made to replace or act as a missing biological structure.

LED: Light Emitting Diode.

Photosensitization: making neurons light sensitive.

Photoreceptor cells: retina cells that convert light into a change of cell membrane potential.

Retina: a very thin, multilayer network of neural cells at the back of the eyeball of vertebrates.

Retinal Ganglion cells: neural cells in the retina which receive the signals from the bipolar and amacrine cells (the cells that process the visual information passed by the photoreceptors) and send action potentials to the brain down axons, which form the optic nerve.

A Novel Blind Wavelet Base Watermarking of ECG Signals on Medical Images Using EZW Algorithm

Mohammad Saleh Nambakhsh
King's College London, UK

M. Shiva
California State University, USA

INTRODUCTION: NECESSITIES, METHODS, AND ALGORITHMS CURRENTLY USED

Exchange of databases between hospitals needs efficient and reliable transmission and storage techniques to cut down the cost of health care. This exchange involves a large amount of vital patient information such as biosignals and medical images. Interleaving one form of data such as 1-D signal over digital images can combine the advantages of data security with efficient memory utilization (Norris, Englehart & Lovely, 2001), but nothing prevents the user from manipulating or copying the decrypted data for illegal uses. Embedding vital information of patients inside their scan images will help physicians make a better diagnosis of a disease. In order to solve these issues, watermark algorithms have been proposed as a way to complement the encryption processes and provide some tools to track the retransmission and manipulation of multimedia contents (Barni, Podilchuk, Bartolini & Delp, 2001; Vallabha, 2003). A watermarking system is based on an imperceptible insertion of a watermark (a signal) in an image. This technique is adapted here for interleaving graphical ECG signals within medical images to reduce storage and transmission overheads as well as helping for computer-aided diagnostics system. In this chapter, we present a new wavelet-based watermarking method combined with the EZW coder. The principle is to replace significant wavelet coefficients of ECG signals by the corresponding significant wavelet coefficients belonging to the host image, which is much bigger in size than the mark signal. This chapter presents a brief introduction to watermarking and the EZW coder that acts as a platform for our watermarking algorithm.

EZW

The EZW algorithm was originally developed by Shapiro (1993) to find the best transmission order of the wavelet coefficients, which is the absolute value of decreasing order. This algorithm has already been applied to medical images and the electrocardiogram with good success (Nambakhsh, Ahmadian, Ghavami & Dilmaghani, 2006).

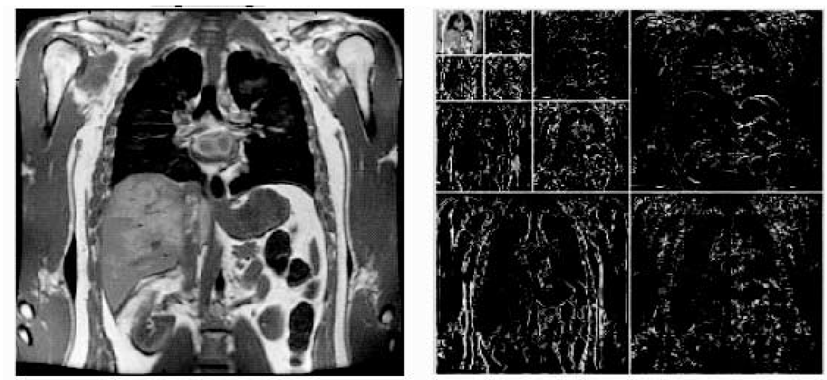
EZW in Image

The wavelet transform is a dyadic decomposition of an image (Nambakhsh et al., 2006) achieved by a pair of quadratic mirror filters (QMF). In two-dimensional separable dyadic discrete wavelet transform (DWT), each level of decomposition produces four bands of data, one corresponding to the low pass band (LL), and the other three corresponding to horizontal (HL), vertical (LH), and diagonal (HH) high pass bands (Figure 1).

Two distinct properties of the EZW algorithm make it an effective means of compression, as compared to traditional approaches. First, the EZW algorithm exploits the hierarchy of the wavelet coefficients and establishes a connection between coefficients from corresponding subbands in different levels of wavelet decomposition, allowing multiple coefficients to be encoded simultaneously. Second, coefficients are encoded in order of importance using bit prioritization (Mallat, 1989). After performing the 2-D DWT on the image, the resulting wavelet coefficients are coded by using a decreasing sequence of thresholds; that is, $T_0, T_1, \ldots, T_{N-1}$, where:

$$T_i = \frac{T_{i-1}}{2}, \qquad T_0 = 2^{\log_2(\max|\gamma(x,y)|)} \qquad (1)$$

Figure 1. Levels of decomposition



Here $\gamma(x,y)$ is the amplitude of wavelet coefficients and T_0 is the initial threshold value. The algorithm executes recursively two successive passes by considering significant coefficients in each pass related to the current threshold only (i.e., absolute value is higher than current threshold).

In the first pass, called the dominant pass, we look through significant coefficients related to the current threshold according to a scan order given in Figure 2(a) using the hierarchy given in Figure 2(b).

The algorithm then provides positions and signs of the significant coefficients in a predefined path that associates the absolute value of the parent coefficients with respect to their children ones. Two kinds of predefined paths are depicted in Figure 3.

Coefficients according to equation 1 and the way of scan that is illustrated in Figure 3 turn into four different symbols in the every step. In other words, they will be coded to P, N, Z, and IZ by the every pass process. Table 1 shows complete instructions of the coding process.

EZW in Signal

The EZW algorithm can be applied on signals similar to image. In this case, decomposed signal coefficients have dyadic tree instead of quad tree (Barni et al., 2001). This is illustrated in Figure 4.

It is obvious that each coefficient has two correlated coefficients in lower scale and so forth. We will use this similarity between images and signals to embed a signal inside an image, which is the topic of the next section.

PROPOSED WATERMARKING ALGORITHM

The multiresolution watermarking with multi secret key algorithm, which is developed in this chapter, is blind, because during the EZW decoding, only the multiparameter's secret key, which includes the header information and many parameters defined by the user, is used to extract the embedded information. There are

Figure 2(a). Scan order of wavelet coefficients (b) quad-tree relation of coefficients

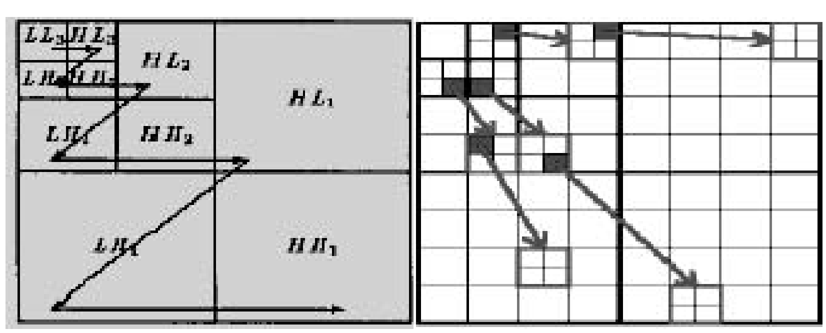


Figure 3. Two kinds of scans of wavelet matrix utilizing in EZW algorithm

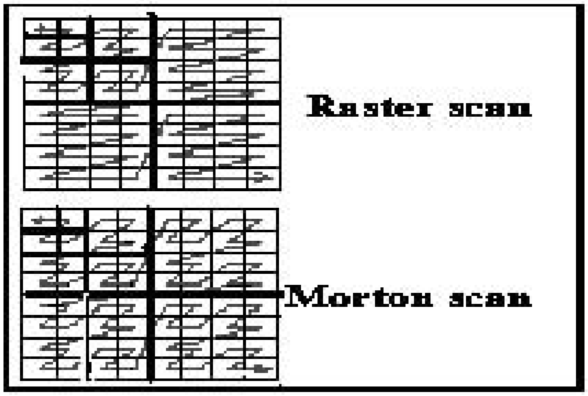
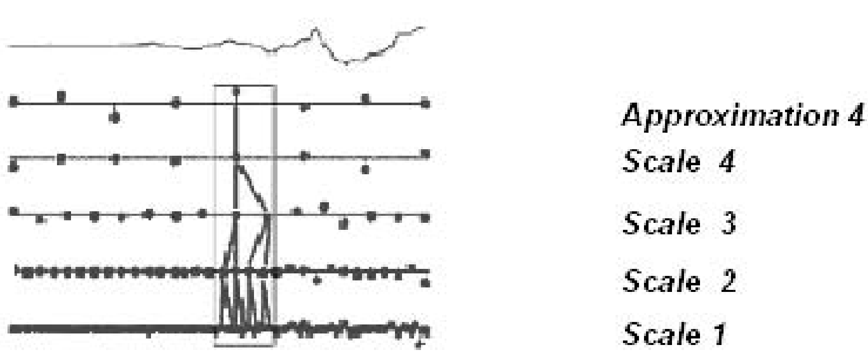


Table 1. Symbol generation instructions

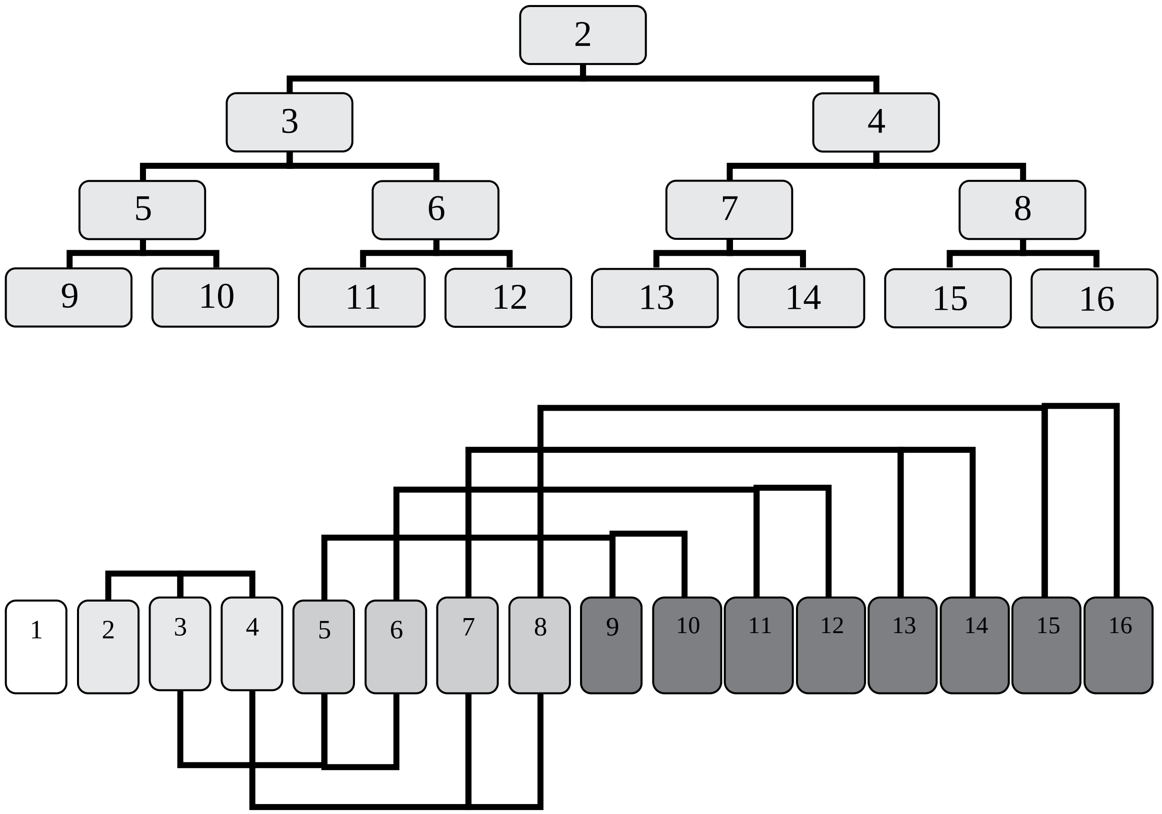
Symbol	Absolute Value	Sign
P Positive & Significant	**Bigger than threshold**	+
N Negative & Significant	**Smaller than threshold**	-
Z Zero-tree	**Smaller than threshold but significant tree include coefficients bigger than threshold**	*
IZ Isolated Zero-tre**e**	**Smaller than threshold and all coefficients of tree are smaller than threshold**	*

Figure 4. Dyadic tree of a decomposed signal



continued on following page

Figure 4. continued



two processes for embedding the mark signal during EZW coder:

The first is insertion process, and the second is extraction process, which is presented in sections 1 and 2.

The following notations will be used:

- INT_i : Thresholds interval $[T_i, T_{i+1}]$
- T_{sig} : Maximum threshold of matrix of signal decomposed signal
- T_{imag}: Maximum threshold of matrix of image decomposed image

Watermark Insertion Process

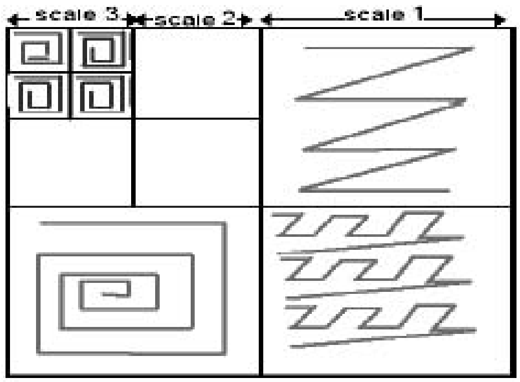
We performed up to five level of wavelet decomposition using Db2 (Wang, Doherty & Van Dyck, 2002) as mother wavelet to obtain matrixes of wavelet coefficients of host image and signal. Then, we specify maximum threshold of both matrixes (Eq.1). In the encoding process of the EZW coder, the image starts to be coded. After this, we look for the dominant pass with a threshold the same as the maximum threshold of the signal (T_{sig}).

After this, we look for the dominant pass with a threshold the same as the maximum threshold of the signal (Tsig). At the beginning, those belonging to this interval (INTi) are replaced with similar coefficients belonging to the last scale of the image matrix in current interval (INTi) according to a predefined path (scan order) of the matrix (Figure 2).

This means that once the coefficients of D4 are replaced with coefficients of image at scale 4 (Figure 6), then the rest of replacement at lower scales can be done by using the child-parent tree (Figures 2 and 4). Note that this predefined path can be changed by the user. In order to better comprehend, Figure 5 shows three different paths.

Subsequently, subtrees (coefficients in the lower scale according to Figure 4) of the embedded signal coefficients are replaced with subtrees (see Figure 2(b)) of image coefficients in a lower scale. Figure 6 illustrates replacement in the algorithm in a general way.

Figure 5. Three predefined paths of embedding

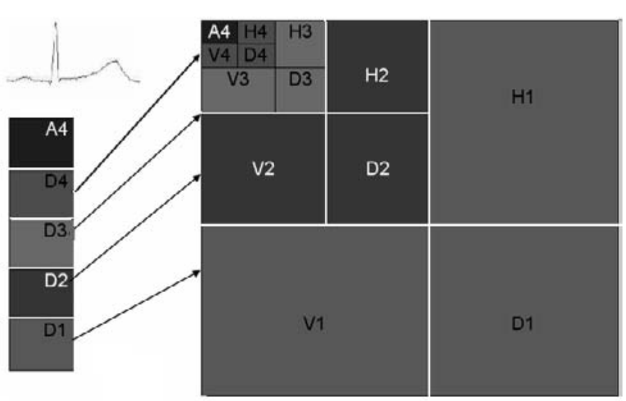


According to Human Visual System (HVS), coefficients located in lower scales are less sensitive than those in higher scales and details (H, V, D) are also less sensitive than approximation (A). Therefore, more significant coefficients are embedded inside a less sensitive region of an image.

HVS has at least two advantages: First, there must be a minimum amount of contrast between the areas of the image to be perceived as different by human eyes. Second, the HVS is less sensitive around edges than in smooth areas of an image. This effect is referred to as "spatial masking" and can be exploited for watermarking by increasing the watermark energy locally near the edge (Voloshynovskiy, Herrigel & Pun, 2000).

As part of the implementation of the algorithm, we use the perceptual model for varying the watermark based on the host image content (Noore, Tungala & Houck, 2004; Podilchuk & Zeng, 1998). An amplifying factor array, α, is computed for different scales and subbands according to human visual system (HVS). In other words, α is an array of amplifying factors for different subbands in different scales. It varies the watermarks coefficients such that a maximum amount of information can be hidden in the host image depending on luminance and contrast properties of the embedding region with least distortion to host image. In other words, the correlation between the original watermark, w_m, and the embedded watermark, w_a, is made maximum, while keeping the perceptual distance between the original image and watermarked image constant. The best value of α is found by iteratively computing the just noticeable difference (JND) for the watermarked image and reducing this difference to a target value. The embedded watermark is finally obtained using the best value defined in Eq. (2):

Figure 6. Correspond scales of replacement



$$Wa_s^d = \alpha_s^d W_{ms}^d \quad (2)$$

Where d is detail subbands that d={1,2,3} is equal with {H,V,D} and s is scale. For example, $\alpha_3^2 = V_3$, Wa_3^2 is watermark stream which would be embedded into V_3.

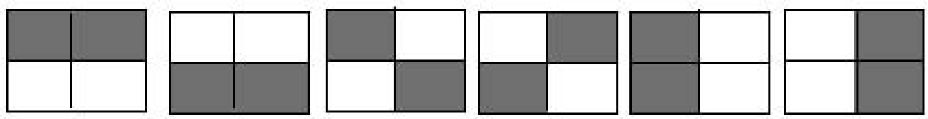
It is important to know that α for the last scale is equal, first because the extraction process has to be able to find coefficients of last scale of signal automatically. Accordingly, we have blind watermarking.

We perform the following process, which tries to insert the wavelet coefficients of the watermark:

1. **Initialization:** A multisecret key is used to select the embedding locations parameterized and randomly to secure the original watermark from tampering. Specifying the multisecret key includes:
    a. Way of scanning in EZW coder; for instance, Morton scan
    b. Prioritization of detail subbands of image for embedding (H, V, D, or V, D, H, ...)
    c. Selecting one way of embedding dyadic tree of the signal inside a quad tree of the image out of six ways (Figure 7)
    d. Maximum level of decomposition for signal and image
    e. Size of the signal
    f. Maximum threshold of the signal
    g. Way of selecting coefficients of signal for embedding in every pass

Figure 8. Block diagram of watermarking embedding process

Figure 7. Six ways of replacing two signals' coefficient with four possible images' coefficient



h. Coordinate of start point of embedding in wavelet matrix of image
i. Definition of mother wavelet.
j. Matrix factor α_S^d in Eq.2

2. **Encoding:** Making list of mixed symbols P, N, Z, IZ:
   i=0;
   $T_{imag} = T_{MAX-of-iaage}$
   $T_{sig} = T_{MAX-of-Signal}$
   $T_i = T_{imag}$

   While i>=0
   $INT = [T_i - T_{i-1}[$
   Check Threshold (Ti);
   If $T_i \leq T_{sig}$
   Find$\{S_i\} \in INT$ in signal detail last scale (n=number of coefficients);
   Find $\{C_i\} \in INT_i$ in image details last scale (H,V,D);
   Replace S_i with C_i;
   Replace sub-tree of S_i inside sub-tree of C_i;
   End;
   EZW (Dominant pass & subordinant pass);
   i=i+1;
   $T_i = T_{i-1} / 2$;
   End;

   This process is nearly similar for the embedding of signal approximation within the image approximation.

3. **Decoding:** The final watermarked image is obtained when the embedded subbands are reconstructed by passing EZW decoders, which turn symbols into wavelet coefficients by refining algorithm with the help of multisecret key information and eventually a five-level inverse discrete wavelet transform (IDWT).

The block diagram of embedding process is illustrated in Figure 8 to clarify different aspect of this section thoroughly.
According to the symbols generated during the encoding process in EZW algorithm, we present a formula for the first time, which represent percentages of redundant data inside images. It will be a comparing measure for considering effects of watermarking in contrast with mark size.

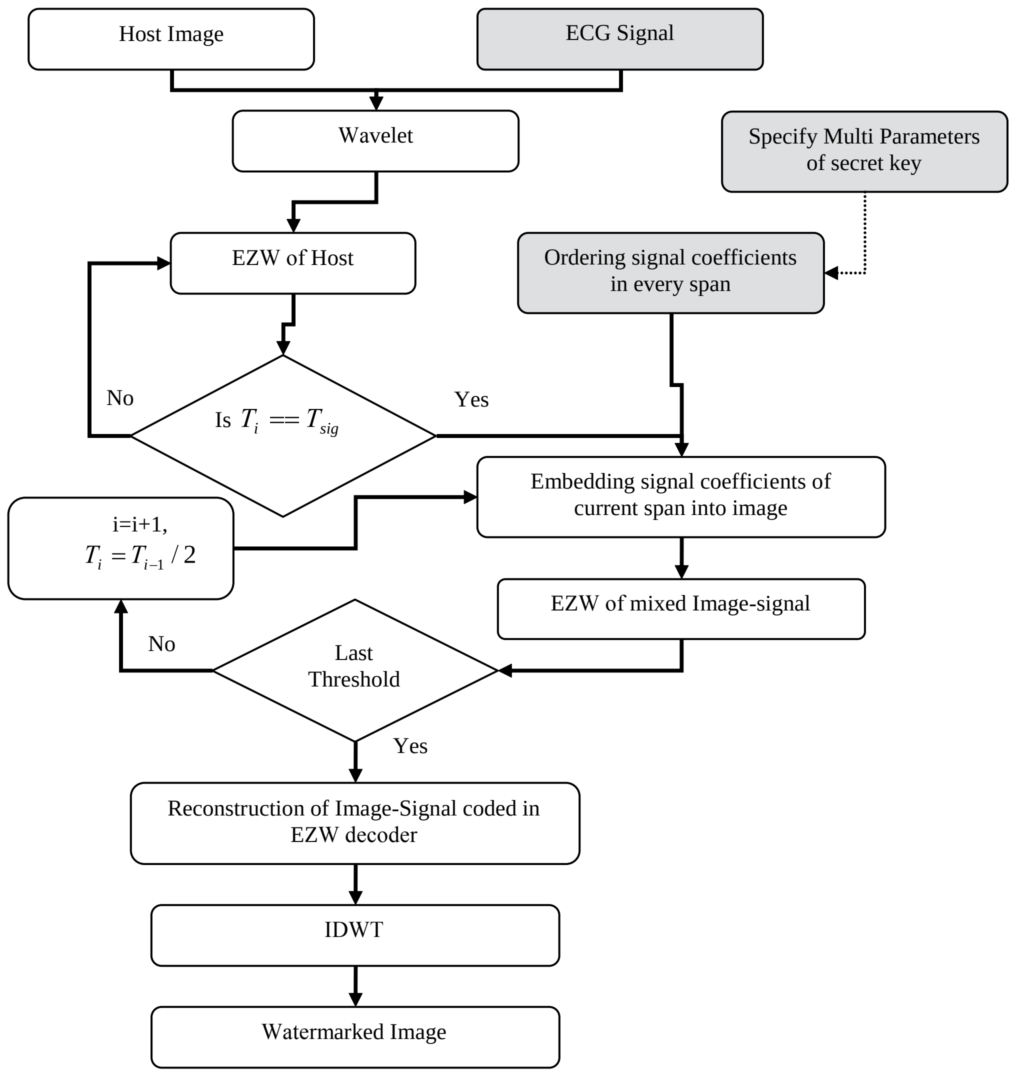
$$RD = \frac{Z+T}{Z+T+P+N} \times 100 \qquad (3)$$

while Z,T,P,N are numbers of symbols that have been defined in Table 1. RD is the percentage of redundancy for each image.

Watermark Extraction Process

In the extraction process, we nearly repeat most of the steps in the insertion process but in the reverse order, and the wavelet coefficients of the compressed watermarked image are constructed. Then, scanning is performed on the matrix in a predefined way using the initial multiparameters' secret key to find signal coefficients. In this step, we only need the wavelet coefficients matrix of the watermarked compressed image to construct the extracted wavelet coefficients of the ECG signal. By using the multiparameters' secret key highest details matrix coefficients (in Figure 6, D4 will be constructed from H4, V4, D4 by the secret key) and then according to their position, other coefficients in the lowest scale will be extracted. Finally, matrix of wavelet coefficients of the ECG will be constructed, and then 1-D and 2-D inverse discrete wavelet transform (IDWT) is applied to the matrix in order to reconstruct the extracted watermark and the compressed watermarked image, respectively. Block diagram of the extraction process has been illustrated to better observation of the process (Figure 9).

Figure 8. Block diagram of watermarking embedding process



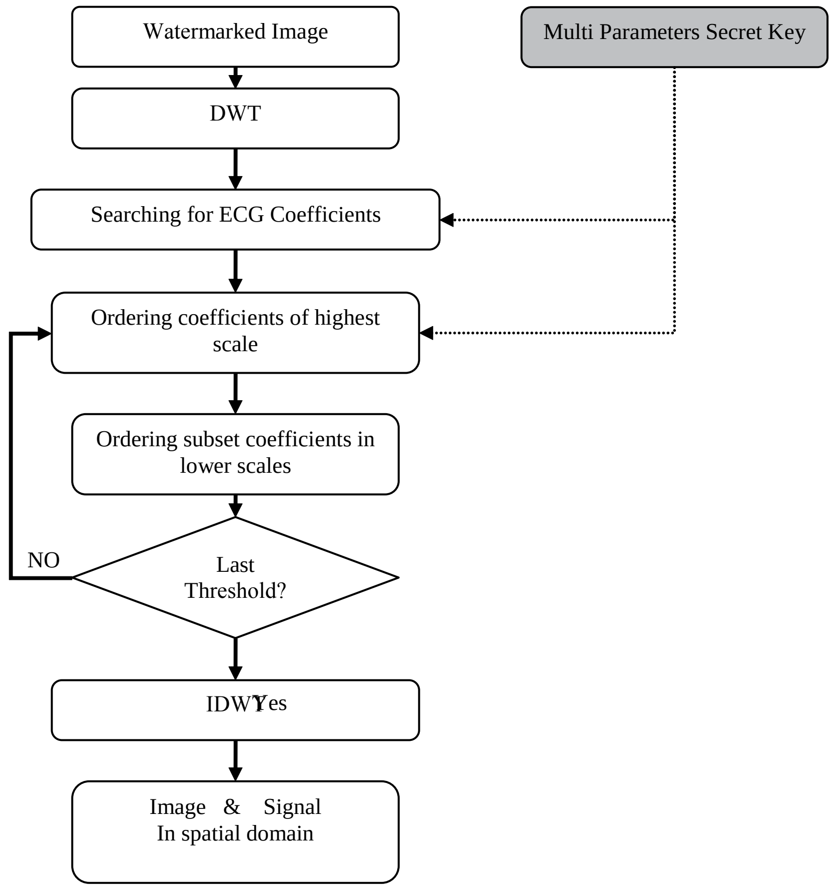
EXPERIMENTAL RESULTS

Several grey-scale MRI and CT images of size 256 × 256 and 512 × 512 pixels have been watermarked by embedding two different watermarks of size 512 and 1024 bytes (Figure 10).

As it is shown, the watermarked images are perceptually or visually acceptable. The original mark ECG signal (1KB) together with its reconstruction after extraction are shown in Figure 11.

We have used several mother wavelets such as Db2 to Db6 and different Cohen filters and found out that Db2 and Db6 are the best for our experiments (Wang et al., 2002). The PSNR calculated between the original image and the compressed watermarked and original signal and extracted one is more than 35 dB for all experi-

Figure 9. Block diagram of extraction process



ments as shown in Figure 12 and Table. 3. This proves the imperceptibility of the embedded watermarks, the property of unchanged remaining extracted signal, and the ability to find the signal accurately.

Beta factor (β) is for evaluation of the degradation of details or edges throughout the image. A β close to one shows less degradation. We have calculated β of the watermarked image as shown in Figure 13 and Table 3 using the following equation:

where ΓI is the high-pass Laplacian filter of the original image, $\overline{\Gamma I}$ is its mean value and wI is the watermarked image.

To compare results of β with amount of RD, we have depicted Table 2 to show the relation of RD with variance (σ^2) of one of the details in scale4 in every image. Obviously, it has inverse relation with variance and direct relation with β, which is resulted after watermarking.

Note that variance in other subbands and scales have direct relation with variance of H4. Therefore, we only show the variance of H4.

From the Figure 14, it is obvious that increase of redundant data in images will provide more space for embedding so that changes in edges are less sensitive.

The proposed method of selecting the insertion sites ensures both watermarking constraints: imperceptibility, since the watermark coefficients of highest scale are almost identical or even equal to those with which they are replaced; robustness since the substituted

Figure 10. (a) original (b) watermarked (ECG size =512) (c) Watermarked (ECG size=1024) images

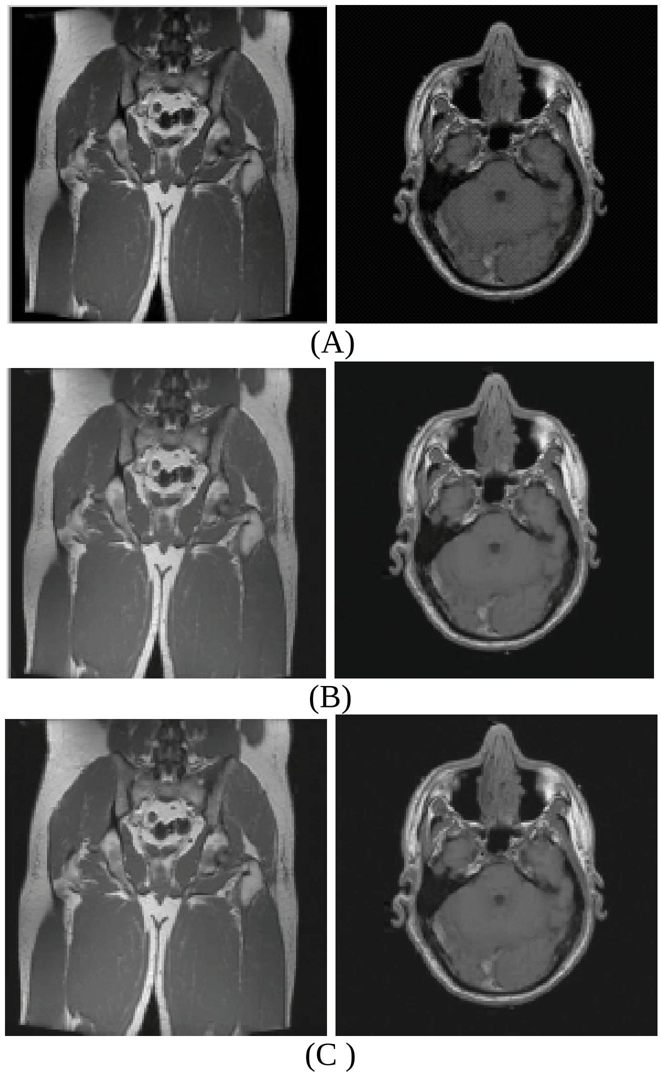


Figure 11. Original signal (continues lines) and extracted signal ('+')

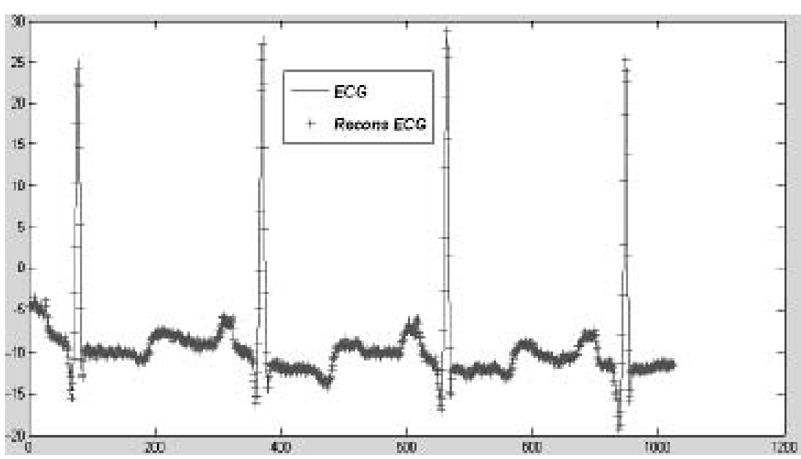


Figure 12. Variation of PSNR vs. mark size for different images

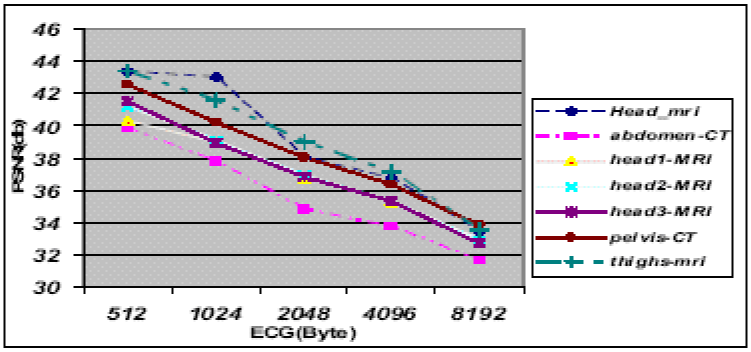


Figure 13. Variation of β vs. mark size for different images

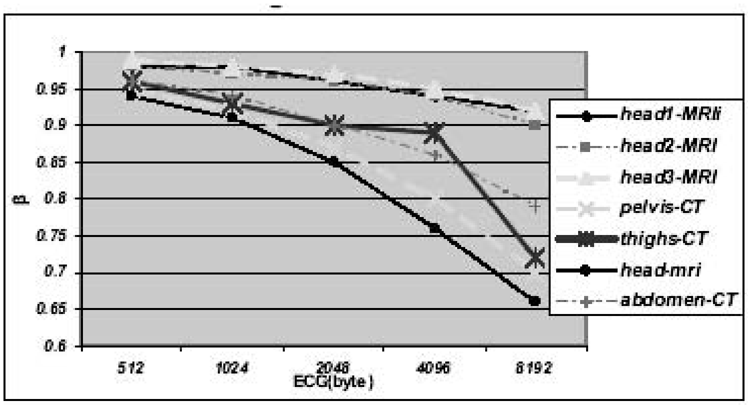


Table 2. Resulting parameters of the RD vs. β and σ^2

Image (EZW)	Size (signal)	Beta (β)	RD (%)	(σ^2)H4)
Head_MRI	512	0.94	69	19614
	1024	0.91	69	19614
017_MRI	512	0.99	86	3953
	1024	0.98	86	3953
Thighs_MRI	512	0.95	76	20464
	1024	0.93	76	20464
Ct_brain1	512	0.98	85	3129
	1024	0.97	85	3129
Ct_brain2	512	0.97	84	3539
	1024	0.96	84	3539
Ct_brain3	512	0.97	84	3407
	1024	0.6	84	3407
CT-VM	512	0.96	80	17050
	1024	0.94	80	17050

Figure 14. Variation of β vs. RD for different images

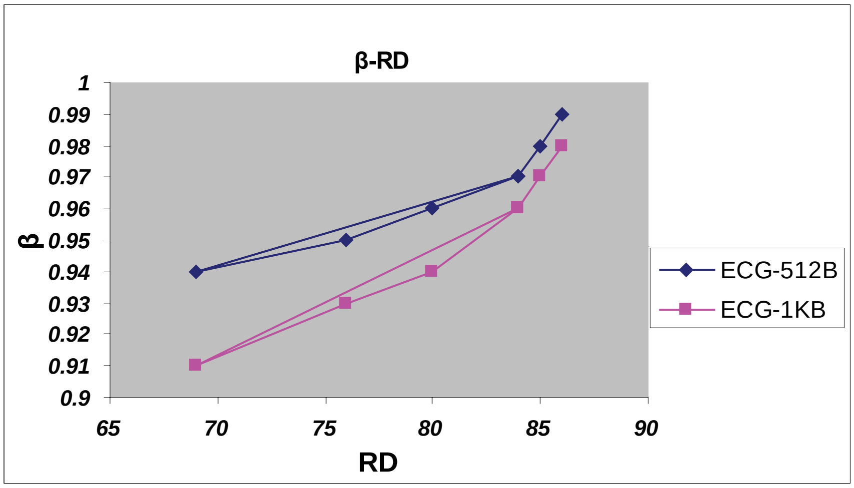


Table 3. Resulting parameters of the numerical simulations

Image	Image Size	Signal (Byte)	Signal NMSE	Image PSNR (db)	Signal PSNR (db)	CC % Signal	B	S/I (%)
Head_MRI	256×256	512	.0008	43.47	58.59	99.88	0.94	.9
	256×256	1024	.0094	42.96	53.66	99.56	0.91	1.8
	256×256	2048	.0317	38.01	42.39	93.97	0.85	3.6
	256×256	4096	.0375	36.75	41.78	92.38	0.76	7.2
	256×256	8192	0.069	33.45	39.05	87.57	0.66	14.4
Head_MrI1	256×256	512	.0008	40.29	58.59	99.88	0.98	.9
	256×256	1024	.0005	38.96	59.79	99.9	0.98	1.8
	256×256	2048	.0005	36.78	59.85	99.90	0.96	3.6
	256×256	4096	.0006	35.25	59.24	99.87	0.94	7.2
	256×256	8192	.0011	32.85	56.84	99.77	0.92	14.4
bnHeadMRi	512×512	512	.0008	48.36	58.59	99.88	0.89	.2
	512×512	1024	.0005	45.58	59.75	99.90	0.85	.4
	512×512	2048	.0005	43.57	59.85	99.90	0.78	.8
	512×512	4096	.0006	41.85	59.24	99.87	0.68	1.6
	512×512	8192	.0010	39.29	57.16	99.79	0.56	3.2

coefficients are significant, and they will not be lost in the quantization step. Moreover, if the watermark extraction process does not find the considered approximation coefficients in the watermarked image matrix at the location indicated by the key, it will search an equivalent value somewhere else. And blindness because extraction algorithm must find the highest detail and approximate according to the information of multisecret key. Cross correlation (CC%) have been calculated between the original and extracted signals for evaluation of percentage of the unchanged signal. S/I (signal/image) is the percent of signal embedded within the image as shown in Table.3.

CONCLUSION AND FUTURE TRENDS

The proposed watermarking algorithm is blind because it only needs an initial key in the extraction process. We have included two processes in the EZW coder; the first one uses embedding the ECG signal in the host image after the decomposition step. The second process intervenes after the decomposition of the compressed watermarked image in order to extract watermarks.

The proposed algorithm has proved its imperceptibility since the embedding process here means a substitution of original coefficients by the equivalents (in the sense of EZW significance) from the watermark. The great advantage of this method is to decode the host image and the mark signal, progressively. This enables extracting the mark signal in a progressive approach from low to high resolution.

The outcome of this proposed method is to make an optimum balance between the resolution of host image and the size of the mark by controlling the maximum scale to be scanned in EZW algorithm. This presents a novel resolution controlled watermarking. Also it is possible to embed biosignals and demographic texts of a patient inside his or her related medical images to reach high integrity between related medical information of each patient and prevent mismatching of information in medical databases. To optimize the proposed method, we are going to use Hermit coefficients of ECG signal instead of its wavelet coefficients. Thus, one can raise capacity of watermarking with almost a factor of six.

REFERENCES

Barni, M., Podilchuk, C.I., Bartolini, F., & Delp, E.J. (2001). Watermark embedding: Hiding a signal within a cover image. *IEEE Communications Magazine, 39*(8), 102–108.

Mallat, S.G. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transaction on Pattern Analysis and Machine Intelligence, 11*(7), 674–693.

Nambakhsh, M.S., Ahmadian, A., Ghavami, M., & Dilmaghani, R. (2006). A novel blind watermarking of ECG signals on medical images using EZW algorithm. In *Proceedings of the IEEE International Conference on Engineering in Medicine and Biology Society*.

Noore, A., Tungala, N., & Houck, M.M. (2004). Embedding biometric identifiers in 2D barcodes for improved security. *J Comput Secur, 23*(82), 679–686.

Norris, J.A., Englehart, K.B., & Lovely, D.F. (2001). Steady-state and dynamic myoelectric signal compression using embedded zero-tree wavelets. In *Proceedings of the 23rd Annual Conference - IEEE/EMBS,* Istanbul, Turkey.

Podilchuk, C.I., & Zeng, W. (1998). Image-adaptive watermarking using visual models. *IEEE J Sel Areas Commun, 16*(4), 525–539.

Shapiro, J.M. (1993). Embedded image coding using zerotrees of wavelet coefficients. *IEEE Transactions of Signal Processing, 41*(12), 3445–3462.

Vallabha, V.H. (2003). *Multiresolution watermark based on wavelet transform for digital images*. Cranes Software International Limited.

Voloshynovskiy, S., Herrigel, A., & Pun, N.B. (2000). A stochastic approach to content adaptive digital image watermarking. *Lecture Notes in Computer Science, 1768*, 212–236.

Wang, Y., Doherty, J.F., & Van Dyck, R.E. (2002). A wavelet-based watermarking algorithm for ownership verification of digital images. *IEEE Transaction Image Processing, 11*(2), 77–88.

KEY TERMS

ECG: Electrocardiograph signals that are recorded from heart activation.

Embedding: Fixing information in other information firmly and deeply.

EZW: Embedded zero-tree wavelet.

Medical Image: Digital images from different parts of the body that are acquired by MRI, CT, Mammography, and so forth; their role here is host in the watermarking process.

Watermarking: Inserting or embedding a kind of digital information inside other kinds or similar kinds of digital information.

Wavelet: A kind of transform domain.

A Novel Radiotherapy Technique

Maria F. Chan
Memorial Sloan-Kettering Cancer Center, USA

Yulin Song
Memorial Sloan-Kettering Cancer Center, USA

INTRODUCTION

Cancer is the second leading cause of death behind heart disease in the United States (National Center for Health Statistics, 2006). Various caner treatments are available now, but the three most common types are surgery, chemotherapy, and radiotherapy. Radiotherapy uses radiation—in the form of a special kind of x-rays, γ-rays, electrons, or protons to kill cancer cells. The advance of modern radiotherapy is closely linked to the evolution of computer and information technology. Intensity-modulated radiation therapy (IMRT) is an advanced form of three-dimensional conformal radiotherapy (3DCRT). It has become today's standard for state-of-the-art radiation treatment. IMRT is based on the concept of inverse treatment planning. By optimizing a constrained objective function, IMRT can modulate the shape and intensity of radiation beams to create a dose distribution highly conformal to the target volume. In this article, we report the results of our investigation on the feasibility and potential benefits of combining electrons with intensity-modulated photons, called IMRT+e, for selected types of caner, particularly for superficial tumors with critical organs underneath. The aim is to deliver high radiation dose to the tumor while sparing all the surrounding normal tissues, thereby minimizing radiation induced side-effects.

BACKGROUND

IMRT uses nonuniform beam intensities within a radiation field to provide more degrees of freedom for dose shaping and dose escalation, resulting in dose distributions that conform more tightly to the tumor shape. As a novel treatment modality, IMRT consists of three major components: a CT-based simulation system, an inverse treatment planning system with computerized optimization, and a computer-controlled linear accelerator delivery system. Unlike traditional IMRT, IMRT+e incorporates static electron fields into photon IMRT fields, taking advantage of electron beams' shallow penetration depth. The novel feature of this approach is that the optimization takes the exiting electron dose distributions into account, resulting in a treatment plan superior to a composite plan whose component plans are optimized independently. Our earlier study (Chan, 2006) has quantified the potential benefits of IMRT+e using established dosimetric parameters for malignant pleural mesothelioma (MPM). Details of various radiation techniques for MPM, such as photon and electron matching (Kutcher, Kestler, Greenblatt, Brenner, Hilaris, & Nori, 1987; Yajnik, Rosenzweig, Mychalczak et al., 2003), intraoperative radiotherapy (IORT) (Hilaris, Nori, Kwong, Kutcher, & Martini, 1984; Lee, Everett, Shu et al., 2002; Rosenzweig, Fax, Zelefsky, Raben, Harrison, & Rusch, 2005), IMRT (Ahamad, Stevens, Smythe et al., 2003; Baldini, 2004; Forster, Smythe, Starkschall et al., 2003; Munter, Christian et al., 2005; Munter, Nill, Thilmann et al., 2003), and intensity-modulated arcs (Tobler, Watson, & Leavitt, 2002), have been published. The first three techniques have been implemented clinically. As demonstrated by our data (Chan, 2006), IMRT+e provided superior target dose coverage compared to the conventional techniques. In addition, IMRT+e also improved critical organ sparing. Here, we present the results of IMRT+e for the treatment of various superficial tumors.

To further explore its potential benefits, IMRT+e was used on a patient with extensive lesions of the scalp covering the entire forehead and both temporal surfaces (Chan, 2006). Traditionally, these lesions have been treated with various electron beams. The relatively high skin dose provided by the electron beams, as well as the limited particle range, allows for adequate dose to the superficial target volume while sparing the underlying normal tissue, principally the brain. To eliminate the field-matching problem with electrons for extensive lesions, several investigations have been attempted by adding photon fields (Akazawa, 1989; Tung, Shiu, Starkschall et al., 1993). Electron arc techniques can also be used although designing an arc to deliver a sufficient dose to the target is difficult. The use of arcing photon fields was also reported (Kinard, Zwicker, Schmidt-Ullrich et al., 1996). Energy- and intensity-modulated electron beams have been investigated (Bedford, Childs, Hansen et al., 2005; Karlsson, Karlsson, & Zackrisson, 1998; Klein, 1998; Locke, Low, Grigireit et al., 2002; Ma, Pawlicki, Lee et al., 2000; Yaparpalvi, Rontenla, & Beitler, 2002). IMRT is able to modulate the photon fields, thereby customizing the dose distribution more specifically to the target. Locke et al. (2002) evaluated the use of tomotherapy with the Peacock system (NOMOS Corporation, Sewickley, PA) for this type of problem. It was found that IMRT provided improved dose homogeneity in the planning target volume (PTV) compared with the more conventional techniques, but resulted in higher doses to the brain and lens. More recently Bedford et al. (2005) further examined the possibility of using IMRT for extensive scalp lesions. They compared photon IMRT with either two or four matched electron fields and arcing electron fields. Their results showed that IMRT offered a feasible alternative to electron techniques for the treatment of extensive scalp lesions. Although it was considered to be clinically acceptable, the brain dose was higher with photon IMRT. In our study, the IMRT+e plan produced a more desirable dose distribution in terms of the dose reduction to the brain with the same dose conformity and homogeneity in the target volumes.

We also employed the IMRT+e technique to treat a patient with Merkel cell carcinoma of the left upper eyelid. A complex IMRT+e plan was computed to cover the left upper eyelid and the left lateral periorbital soft tissues, as well as the facial, periparotid, and cervical lymphatics. The plan consisted of three 6 MV IMRT fields and one enface 6 MeV electron field, delivering a total dose of 54 Gy in 30 fractions.

IMRT+e METHODOLOGY

The key concept of IMRT+e is to incorporate existing static electron fields into the photon beam inverse planning optimization. The goals are to improve the target dose conformity, boost the skin dose, and spare the distant critical organs. Although the photon beam IMRT has been shown to be an effective treatment for deep-seeded tumors, it is not suitable for treating very shallow targets like skin cancer, certain types of MPM, and breast cancer. This is due to the low surface dose and deep photon beam penetration. The slow attenuation of photon beams can still deliver a significant dose to the distant critical organs. The rapid dose falloff of electron beams makes electron therapy a viable treatment modality for these shallow targets. In addition, electron beams have negligible scatter radiation, compared to photon beams. Furthermore, electron beam therapy uses normal incidence; it is, therefore, less affected by patient's respiration than photon beam IMRT, particularly beamlet-based IMRT.

Prior to optimization, the PTV is delineated on CT images by a radiation oncologist. An appropriate margin is added to the PTV to account for patient setup uncertainty and internal organ movement. Critical structures are also delineated by either a medical physicist or a radiation oncologist. The IMRT+e plan is computed on a planning system developed at the Memorial Sloan-Kettering Cancer Center. The photon-only IMRT plan is computed first using six MV photons with several beam directions. Depending on disease sites, the beam configuration varies with the target and critical organ geometry. To create an IMRT+e plan, one or two electron fields using 6-16 MeV are added to the computed photon IMRT plan. The IMRT plan is then optimized again. This new round of optimization takes the existing electron dose distributions into account. Therefore, for those voxels with exiting electron dose, the photon IMRT fields will deposit less of a dose there. Normally, enface electron field arrangement is sufficient to achieve the acceptable tolerance levels for critical structures. When these acceptable levels can not be met, alteration of electron energy or differential weighting of the electron plan is attempted. In general, the electron fields are manually optimized to contribute about 50% of the prescribed dose. The quality of the final plan is evaluated using dose-volume histograms (DVHs). The photon IMRT component of the plan is designed to be delivered using the dynamic multileaf collimator (DMLC) technique while the static electron component is treated with custom electron cutouts.

The IMRT optimization algorithm uses an iterative gradient search method to minimize an objective function (Chui & Spirou, 2001; Spirou & Chui, 1998). The objective function consists of terms corresponding to the targets and the organs at risk (OAR).

$$F = F_{t\arg e-1} + F_{t\arg e-2} + \cdots + F_{OAR-1} + F_{OAR-2} + \cdots \quad (1)$$

The term for the target is given by:

$$F_{\text{target}} = \frac{1}{N_t}\left[\begin{array}{l}\sum_{i=1}^{N_t}(D_i - D_{presc})^2 \\ +w_{t,\min}\cdot\sum_{i=1}^{N_t}(D_i - D_{\min})^2\cdot\Theta(D_{\min} - D_i) \\ +w_{t,\max}\cdot\sum_{i=1}^{N_t}(D_i - D_{\max})^2\cdot\Theta(D_i - D_{\max})\end{array}\right] \quad (2)$$

where N_t is the number of points in the target, D_i is the dose to point *i*, and D_{presc} is the prescription dose. The second and third terms inside the brackets implement the target dose homogeneity criterion: D_{min} and D_{max} are the desired minimum and maximum target doses, and $w_{t,min}$ and $w_{t,max}$ are the penalties associated with under- and overdosing. Θ(x) is the Heaviside function, defined as:

$$\Theta(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (3)$$

Similarly, the term for the OAR is given by:

$$F_{OAR} = \frac{1}{N_{OAR}}\left[\begin{array}{l} w_{OAR,\max}\cdot\sum_{i=1}^{N_{OAR}}(D_i - D_{\max})^2\cdot\Theta(D_i - D_{\max}) \\ +w_{OAR,dv}\cdot\sum_{i=1}^{N_{dv}}(D_i - D_{dv})^2\cdot\Theta(D_i - D_{dv})\end{array}\right] \quad (4)$$

where the first term inside the brackets implements a maximum dose constraint D_{max} on the OAR and the second term implements a dose-volume constraint. The relative penalty weights are given by $w_{OAR,max}$ and $w_{OAR,dv}$, respectively. N_{OAR} is the number of points in the OAR and N_{dv} the number of points whose dose must be below the dose-volume constraint dose D_{dv}.

The dose to point *i*, D_i, is the sum of dose contributions from all rays:

$$D_i = \sum_{j=1}^{N_r} x_j a_j \quad (5)$$

where N_r is the number of rays, x_j is the intensity of the *j*-th ray, a_{ij} the dose deposited to the *i*-th point per unit intensity of the *j*-th ray, and the product is summed over all *j*. The goal of optimization is to find the set of ray intensities x_j that minimizes the objective function given by expression (1).

The described methodology is, in principle, applicable not only to intensity-modulated photons, but also electrons or combination of both. At this time, medical linear accelerators are not equipped with electron MLC (EMLC), thus intensity-modulated electron plans can not be delivered. As a result, inverse treatment planning is routinely performed for photon beams only. However, to fully take advantage of the finite range of electron beams, it is beneficial to combine electrons with photon IMRT for certain disease sites. This can be accomplished by optimizing photon IMRT over a manually planned electron dose distribution. In this case, D_i in the above expressions is substituted with $D_i + D_{i,\,electron}$, where $D_{i,\,electron}$ represents the electron dose distribution. $D_{i,\,electron}$ is manually determined by the treatment planner and is not varied by the optimization algorithm. Figure 1 shows a representative photon IMRT field, where a low intensity region corresponding to the electron field was created to spare the kidney in an MPM patient.

Several dosimetric parameters from DVHs were evaluated to assess the significance of improvement of this new technique. The target coverage was measured by the dose covering 95% of the PTV (D_{95}), the volume of the target receiving at least 95% of the prescription dose (V_{95}), and the maximum and minimum target doses. The mean doses were used as the plan evaluation criteria for critical organs. The estimated normal tissue complication probability (NTCP) (Burman, Kutcher, Emami, & Goitein, 1989; Kutcher & Burman, 1991; Lyman, 1985) and fraction of lung subunits damaged (Fdam) (Jackson, Ten Haken, Robertson, Kessler, Kutcher, & Lawrence, 1995) were also used to evaluate the probability of pneumonitis for MPM patients. Thermoluminescent dosimeters (TLD), diodes, Gafchromic EBT films, and optically stimulated luminescence (OSL) dosimeters were used to verify the accuracy of the IMRT+e dosimetry. Figures 2, 3, and 4 show representative dose distributions of three different IMRT+e plans.

Figure 1. A BEV of a photon IMRT field (MLC aperture in green) shows a low intensity region (50% intensity level in white) corresponding to the electron field cutout. Kidneys and liver are in pink and light blue, respectively. The yellow mesh represents the shape of the PTV. The distance between grid points is 2 cm.

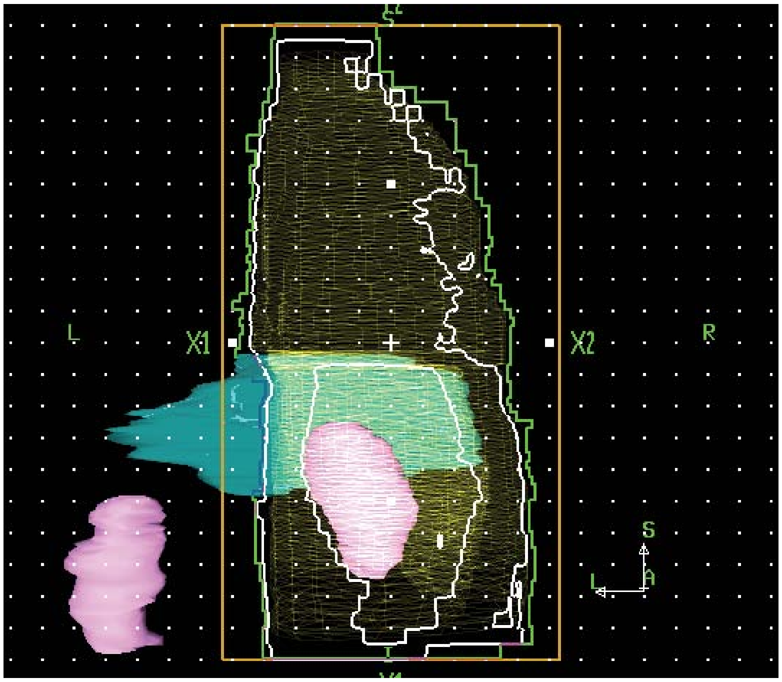


Figure 2. Isodose distributions of a representative transverse image at the level of the kidneys/liver of an MPM patient (PTV in yellow)

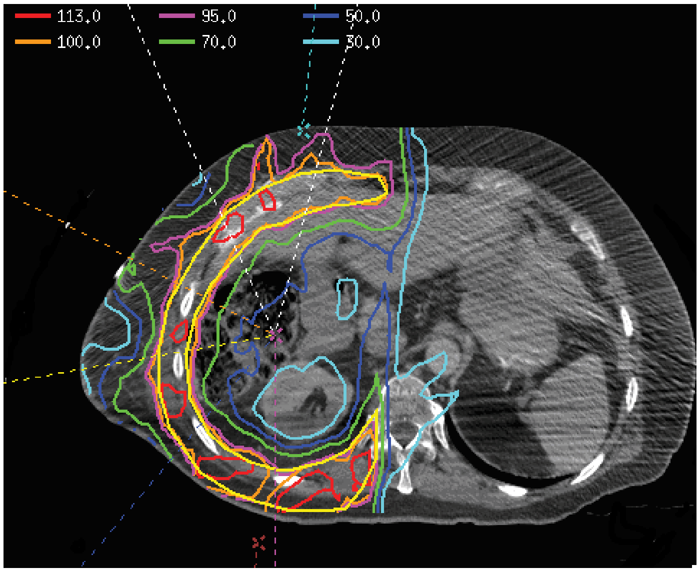


Figure 3. Isodose distributions of a representative sagittal image of the patient with extensive lesions of scalp (PTV in yellow)

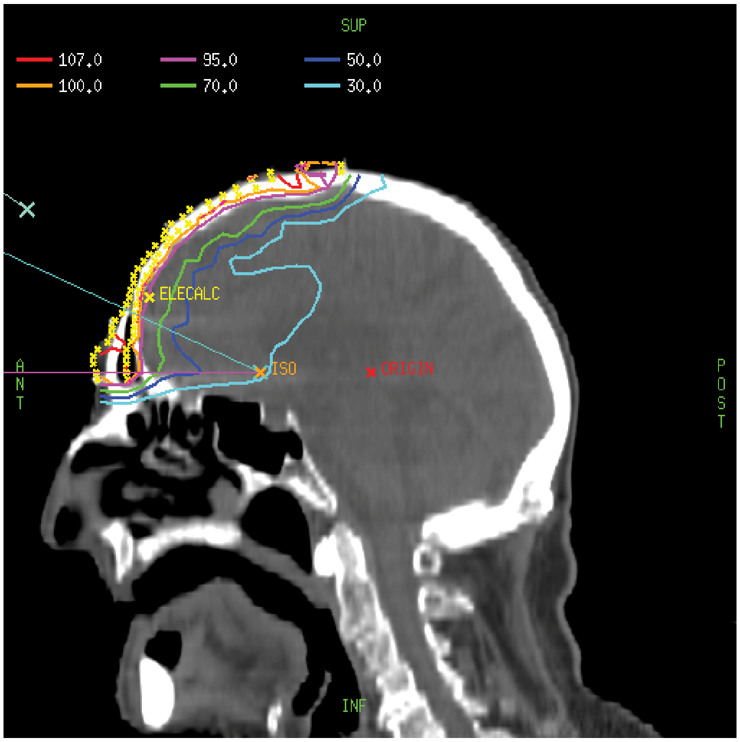
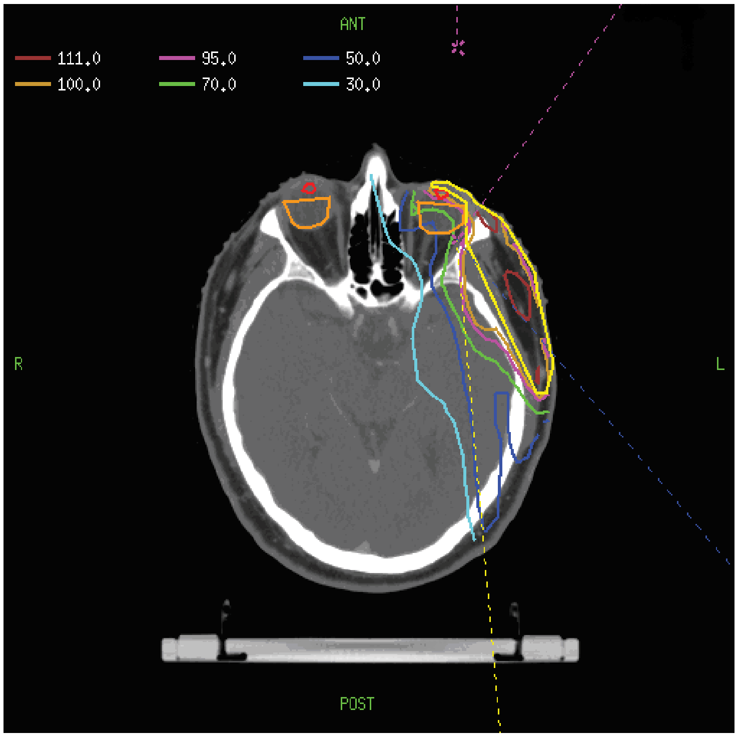


Figure 4. Isodose distributions of a representative transverse image of the patient with Merkel cell carcinoma of the left upper eyelid (PTV in yellow)



FUTURE TRENDS

Recently, there has been a strong interest in using photon IMRT to treat mesothelioma (Ahamad, Stevens, Smythe et al., 2003; Forster et al., 2003; Munter et al., 2003; Munter et al., 2005; Zierhut, Gutwein, Munter, Woger, & Debus, 2004). IMRT has shown great potential to produce a highly conformal dose distribution around the concave-shaped target volumes and a steep dose gradient near the OAR. These characteristics can spare radiosensitive normal structures and reduce complication rates. Therefore, IMRT may improve local-regional tumor control through dose escalation while maintaining acceptable levels of normal tissue complications. However, the high photon exit dose and low surface dose make it suboptimal for treating very shallow targets such as certain portions of typical pleural mesotheliomas.

Extensive scalp lesions also present a challenge to radiotherapy due to the fact that they are superficial and surround critical structures (e.g., brain and optic structures). To search for the best treatment option, a variety of planning techniques has been developed and tested. Bedford et al. (2005) proved that IMRT offers a feasible alternative to electron techniques for the treatment of extensive scalp lesions. The PTV dose distribution with IMRT does not suffer from problems of matching static electron fields and is considerably more homogenous than with either static or arcing electrons. However, the brain dose is higher with IMRT, but it is considered to be clinically acceptable. With IMRT+e, the brain dose can be further reduced, thus minimizing radiation-induced complications.

In theory, tumors with both deep and superficial extensions can be best treated with a combination of photon IMRT and modulated electron radiation therapy (MERT) (Song, Boyer, Pawlicki et al., 2004). Unlike photon IMRT, MERT can provide both intensity and energy modulations. However, the delivery of MERT plans requires an EMLC, which is not available on current commercial medical linear accelerators. Thus, IMRT+e is a promising alternative approach by which a conventionally delivered electron contribution is incorporated into the inverse treatment planning process. The rapid dose falloff of electron beams makes them a preferable treatment modality for the shallow parts of the PTV. Conceptually, such a combined modality plan would consist of multiple photon beams with the same low photon energy and electron beams with different energies, depending on the tumor depth. Dose conformity in the depth direction would be achieved by the use of different electron energies. Dose conformity and uniformity in the lateral directions would be achieved by photon beam intensity modulation using DMLC. Through both intensity and energy modulations, this combined approach is capable of delivering highly conformal doses to targets with complex shapes like mesotheliomas and of improving normal tissue sparing, particularly the distant critical structures. This technique, although not yet clinically mature, might present an opportunity for further dose escalation to the target and could potentially lead to improved local control. Adding the electron beams takes an experienced planner using our planning system approximately 30 minutes longer than IMRT planning alone. Our experience is that IMRT+e treatment time will be 15 minutes longer than IMRT alone.

CONCLUSION

We explored the feasibility of treating mesotheliomas, extensive squamous cell carcinoma of the scalp, and Merkel cell carcinoma of the eyelid using IMRT+e. This approach combined photon IMRT and static electron beams into a single plan. The optimization of the IMRT fields accounted for the existing electron dose contributions. Based on our initial experience, we believe that this new technique is straightforward and clinically viable. It improves target dose coverage compared to conventional non-IMRT methods and increases normal tissue sparing compared to photon-alone IMRT. Further refining of the IMRT+e technique will offer opportunities to escalate tumor dose without increasing normal tissue complication.

REFERENCES

Ahamad, A., Stevens, C.W., Smythe, W.R., et al. (2003a). Intensity-modulated radiation therapy: A novel approach to the management of malignant pleural mesothelioma. *Int J Radiat Oncol Biol Phys., 55*(3), 768-775.

Ahamad, A, Stevens, C.W., Smythe, W.R., et al. (2003b). Promising early local control of malignant pleural mesothelioma following postoperative intensity-modulated radiotherapy (IMRT) to the chest. *Cancer J*., *9*(6), 476-484.

Akazawa, C. (1989). Treatment of the scalp using photon and electron beams. *Med Dosim., 14,* 129-131.

Baldini, E.H. (2004). External beam radiation therapy for the treatment of pleural mesothelioma. *Thorac Surg Clin*., 14(4), 543-548.

Bedford, L., Childs, P.J., Hansen, V.N., et al. (2005). Treatment of extensive scalp lesions with segmental intensity-modulated photon therapy. *Int J Radiat Oncol Biol Phys., 62,* 1549-1558.

Burman, C., Kutcher, G.J., Emami, B., & Goitein, M. (1989). Fitting of normal tissue tolerance data to an analytic function. *Int J Radiat Oncol Biol Phys., 16,* 1623-1630.

Chan, M.F., Chui, C.S., Song, Y., Burman, C., Yorke, E., Della-Biancia, C., et al. (2006). A novel radiation therapy technique for malignant pleural mesothelioma combing electrons with intensity-modulated photons. *Radioth Oncol*., *79*(2), 218-223.

Chan, M.F., Song, Y., Burman, C., Chui, C.S., & Schupak, K. (2006). The treatment of extensive scalp lesions combining electrons with intensity-modulated photons. *IEEE-Engineering in Medicine and Biology,* 152-155.

Chui, C.S., & Spirou, S. (2001). Inverse planning algorithms for external beam radiation therapy. *Med Dosime*., *26,* 189-197.

Forster, K.M., Smythe, W.R., Starkschall, G., et al. (2003). Intensity-modulated radiotherapy following extrapleural pneumonectomy for the treatment of malignant mesothelioma: Clinical implementation. *Int J Radiat Oncol Biol Phys., 55*(3), 606-616.

Hilaris, B.S., Nori, D., Kwong, E., Kutcher, G.J., & Martini, N. (1984). Pleurectomy and intraoperative brachytherapy and post-operative radiation in the treatment of malignant pleural mesothelioma. *Int J Radiat Oncol Biol Phys*., *10*(3), 325-331.

Jackson, A., Ten Haken, R.K., Robertson, J.M., Kessler, M.L., Kutcher, G.J., & Lawrence, T.S. (1995). Analysis of clinical complication data for radiation hepatitis using a parallel architecture model. *Int J Radiat Oncol Biol Phys., 31*(4), 883-891.

Karlsson, M.G., Karlsson, M., & Zackrisson, B. (1998). Intensity-modulation with electrons: Calculations, measurements, and clinical applications. *Phys Med Biol., 43,* 1159-1169.

Kinard, J.D., Zwicker, R.D., Schmidt-Ullrich, R.K., et al. (1996). Short communication: Total craniofacial photon shell technique for radiotherapy of extensive angiosarcomas of the head. *Br J Radiol., 69,* 51-355.

Klein, E.E. (1998). Modulated electron beams using multi-segmented multileaf collimation. *Radiother Oncol., 48***,** 307-311.

Kutcher, G.J., & Burman, C. (1991). Calculation of complication probability factors for nonuniform normal tissue irradiation: The effective volume method. *Int J Radiat Oncol Biol Phys., 21,* 137-146.

Kutcher, G.J., Kestler, C., Greenblatt, D., Brenner, H., Hilaris, B.S., & Nori, D. (1987). Technique for external beam treatment for mesothelioma. *Int J Radiat Oncol Biol Phys*., *13*(11), 1747-1752.

Lee, T.T., Everett, D.L., Shu, H.K., et al. (2002). Radical pleurectomy/decortication and intraoperative radiotherapy followed by conformal radiation with or without chemotherapy for malignant pleural mesothelioma. *J Thorac Cadiovasc Surg*., *124*(6), 1183-1189.

Locke, J., Low, D.A., Grigireit, T., et al. (2002). Potential of tomotherapy for total scalp treatment. *Int J Radiat Oncol Biol Phys., 52,* 553-559.

Lyman, J.T. (1985). Complication probability as assessed from dose-volume histograms. *Radiat Res., 104,* S13-S19.

Ma, C.M., Pawlicki, T., Lee, M.C., et al. (2000). Energy- and intensity-modulated electron beams for radiotherapy. *Phys Med Biol., 45,* 2293-2311.

Munter, M.W., Christian, T., Anna N., Simeon N., & Jurgen D. (2005). Inverse planned stereotactic intensity-modulated radiotherapy (IMRT) in the palliative treatment of malignant mesothelioma of the pleura: The Heidelberg experience. *Lung Cancer, 49S*(1), S83-S86.

Munter, M.W., Nill, S., Thilmann, C., et al. (2003). Stereotactic intensity-modulated radiation therapy (IMRT) and inverse treatment planning for advanced pleural mesothelioma. Feasibility and initial results. *Strahlenther Onkol., 179*(8), 535-541.

National Center for Health Statistics. (2006). Centers for Disease Control and Prevention.

Rosenzweig, K.E., Fax, J.L., Zelefsky, M.J., Raben, A., Harrison, L.B., & Rusch, V.W. (2005). A pilot trial of high-dose-rate intraoperative radiation therapy for malignant pleural mesothelioma. *Brachytherapy, 4*(1), 30-33.

Song, Y., Boyer, A., Pawlicki, T., et al. (2004). *Modulated electron radiation therapy-emerging technology.* In *Intensity-modulated radiation therapy – a clinical perspective.* Lewiston, NY: BC Decker.

Spirou, S.V., & Chui, C.S. (1998). A gradient inverse planning algorithm with dose-volume constraints. *Med Phys., 25,* 321-333.

Tobler, M., Watson, G., & Leavitt, D.D. (2002). Intensity-modulated photon arc therapy for treatment of pleural mesothelioma. *Med Dosime, 27*(4), 255-259.

Tung, S.S., Shiu, A.S., Starkschall, G., et al. (1993). Dosimetric evaluation of total scalp irradiation using a lateral electron-photon technique. *Int J Radiat Oncol Biol Phys., 27,* 153-160.

U.S. Mortality Public Use Data Tape 2003. (2006).

Yajnik, S., Rosenzweig, K.E., Mychalczak, B., et al. (2003). Hemithoracic radiation after extrapleural pneumonectomy for malignant pleural mesothelioma. *Int J Radiat Oncol Biol Phys*., *56*(5), 1319-1326.

Yaparpalvi, R., Fontenla, D.P., & Beitler, J.J. (2002). Improved dose homogeneity in scalp irradiation using a single set-up point and different energy electron beams. *Br J Radiol., 75***,** 670-677.

Zierhut, D., Gutwein, S., Munter, M.W., Woger, H., & Debus, J. (2004). Radiation therapy of mesothelioma: The Heidelberg experience and future aspects. *Lung Cancer, 45S,* S85-91.

KEY TERMS

Three-Dimensional Conformal Radiotherapy (3DCRT): A complex process that begins with the creation of 3D digital data sets of patient tumors, and normal anatomy. These data are then used to generate 3D computer images and to develop complex treatment plans to deliver a highly "conformed" 3D radiation dose while sparing normal adjacent tissues.

IMRT: Intensity-modulated radiation therapy is an advanced form of 3DCRT. It uses sophisticated software and hardware to modulate the shape and intensity of radiation beams targeted on different parts of the treatment area.

Linear Accelerator: The most commonly used device for external beam radiation treatment for cancer patients. It delivers radiation dose of high-energy photons and/or electrons to the region of the patient's tumor.

PTV: Planning target volume is a composite margin of subclinical margin of gross tumor volume, internal margin accounting for variations due to organ movement and respiration, and set-up margin accounting for all the uncertainties in patient/beam positioning.

DMLC: IMRT plans can be delivered with a conventional multileaf collimator (MLC) in the dynamic mode. The MLC leaves continuously move while the beam is on. DMLC delivers intensity-modulated beam profiles previously determined by plan optimization algorithms.

Online Nurse Education

Jane Klobas
Università Bocconi, Italy
University of Western Australia, Australia

Ciro Sementina
TILS SpA, Italy

Stefano Renzi
Università Bocconi, Italy

ABSTRACT

In many countries, healthcare professionals are required to participate annually in compulsory continuing medical education (CME). The effort involved in providing wide-scale training led the Italian Ministry of Health to support pilot courses using online distance learning. This article reports the results of a short survey which aimed to gauge the potential of online CME for nurses in Italy. Most of the 152 respondents, all of whom had completed an online course, supported the inclusion of some form of collaborative learning. Three possible market segments for online learning emerged from the study: nurses who prefer to study alone, those who would appreciate collaborative activities well-integrated into course design, and those who would prefer courses that include online collaboration of any kind. The authors conclude that online learning is a suitable mode for enabling participation in CME for accreditation, but caution that further research is required to confirm that the preferences of nurses who have experienced online distance learning are shared by those who have not.

INTRODUCTION

Rapid developments in medical knowledge have been accompanied by increased attention to continuing medical education (CME) for doctors, nurses, and other healthcare professionals. To ensure practitioners' knowledge remains current, professional organizations or national bodies, or both, typically establish a minimum number of hours or CME credits which a healthcare professional must attain in a year to maintain their accreditation to practice. Given the extent of training required, several national governments have supported pilot studies that use information and communications technologies (ICT) in different ways to enable healthcare professionals to participate in CME from their home or place of work, or while mobile, "on the road."

There is little literature on CME in the information systems field, but some recent studies have examined aspects of interaction by medical workers with Internet and mobile technologies. Studies of adoption of the Internet and Internet-based systems to gather data for evidence-based practice in the United Kingdom in the late 1990s indicated that voluntary use of such systems was limited by the busy day-to-day working life of the staff of medical practices (Abdulrasul, 2001; Howcroft & Mitev, 2000; Klecun & Cornford, 2003). The studies of mobile systems indicate that adoption of mobile technologies in medical contexts is complex, and related at least as much to organizational issues such as control, and extra-organizational issues such as the individual's perceived role as a professional and member of society, as it is to the interface between the user and the technology (Scheepers, Scheepers, & Ngwenyama, 2006; Wiredu & Sorenson, 2006). To this extent, these studies touch on some of the problems that online CME aims to address.

Nurses, like other healthcare workers, are faced with the challenge of combining education with practice, and their lives as professionals, family members, and members of society. In many countries, nurses obtain formal qualifications or keep up to date with developments in theory and practice part-time, juggling study with the demands of their day-to-day work and family life. Since 2002, the Italian government has required health professionals to obtain a minimum number of credits through participation in CME. In 2006, the minimum number of credits, for example, was 50, the equivalent of more than a week of full-time education (Ministero della Salute, 2000). It would be impossible to rely on classroom-based education to implement a nation-wide initiative on this scale. Instead, the Ministry for Health took advantage of the opportunities for distance learning offered by the Internet. In this article, we report on the preferences of nurse participants in one of the pilot courses supported by the Ministry. The course was conducted within the ambit of a research project that sought to evaluate the role of social learning, and in particular, online collaborative learning in motivating participants and reducing dropout rates in online professional development, so we were particularly interested in the participants' preferences for social and collaborative approaches to online learning.

BACKGROUND AND LITERATURE

The Internet permits busy professionals to participate in continuing professional education (CPE) from their offices or homes at times that are convenient to them. Most Internet-based or "online" courses for healthcare practitioners involve delivery of material prepared by professional teachers and trainers in computer-based training packages. These packages typically involve pages or slides that present foundation information and exercises and assessments that enable the learner to practice and test their understanding of the delivered information (Ricketts, Price, & Chamberlain, 2005). Some courses may include multimedia simulations. In many cases, an online tutor is available to answer participant questions and, in some cases, to stimulate involvement and participation in the course (Diekelmann & Mendias, 2005).

One of the problems associated with online courses is that dropout rates are high, even when tutors are available (Parker, 2003). Online social interaction among learners appears to reduce the sense of isolation that is often felt by learners taking courses at a distance from the teacher and other learners (Contreras-Castillo, Favela, Perez-Fragoso, & Santamaria-del-Angel, 2004). Participants in courses that incorporate such online social interaction have been found to have higher satisfaction and higher completion rates than participants in courses that afford no opportunity for social interaction (Renzi & Klobas, 2002). Collaborative learning activities that incorporate social interaction are also believed to improve learners' engagement with the course material and their learning from participation (Rudestam & Schoenholtz-Read, 2002).

A question mark hangs over the value of online social interaction in courses for nurses, however. Platzer, Blake, and Ashford (2000) found that British nurses did not necessarily enjoy group work for CPE in a face-to-face setting. They identified several barriers to learning from group work, including the ways in which the nurses interact with one another and group members' commitment to shared learning. On the other hand, Buckingham (2003) found that Canadian bachelor's degree participants studying at a distance from one another appreciated online discussions, which improved a number of skills, including time management and critical thinking. The differences in the settings of these studies, as well as the differences in the results, make it difficult to conclude if online collaborative learning is a suitable CPE method for nurses, or if nurses who participate in CPE would enjoy such an approach. Indeed, nurse educators have noted the lack of research on Web-based learning, and called for further research in this field (Howatson-Jones, 2004; Kenny, 2000).

In this article, we examine nurses' views of online learning in general, and online collaborative learning in particular. Working with Italian nurses who had just completed an online CME course, we asked two questions:

1. What are the perceived advantages of participation in online professional development courses for nurses?
2. What are nurses' attitudes to participation in courses that involve social and online collaborative learning?

We used the answers to these questions to draw initial conclusions about the nature of online courses that would offer a satisfactory distance learning experience for nurses.

METHODS

The research questions were addressed in a questionnaire survey of nurses who participated in a pilot online CME course supported by the Italian government.

The Course

The course concerned accreditation and measurement of the quality of health services. It was offered by an e-learning provider, TILS SpA, in partnership with the Tor Vergata University Hospital Clinic (Azienda Ospedaliera Universitaria Policlinico Tor Vergata), the Lazzaro Spallanzi National Institute for Infectious Diseases (l'Istituto Nazionale per le Malattie Infettive Lazzaro Spallanzani), and a consulting firm, SI-IES Istituto Europeo Servizi. As a pilot for the national CME initiative, the course was offered free of charge to participants. Participants who successfully completed the course were awarded 12 CME credits (just over 20% of their annual requirement).

The course was delivered entirely online over the World Wide Web (WWW) in the TILS Learning Management System (LMS). Trainees used their web browser to access the course on a site mounted by TILS (http://www.tils.com/ecm). Each of the course's three modules consisted of a lesson presented over a series of screens and supported by a glossary, documents such as copies of relevant directives, and references. The course environment also included an open forum where trainees could discuss the course and course-related issues in general. Trainees could take the course autonomously at any time of day, and could take as much time as they liked to complete each module and the course as a whole, provided they completed the course by the closing date two months from the date that the course was mounted. A tutor was available from within the course environment to answer questions and discuss the course material, although interaction with the tutor was not required to complete the course. Participants also had access to a secretary to help with administrative matters, and a help desk to discuss and resolve problems associated with accessing and using the course technology.

Successful course completion required reaching a minimum standard in a multiple choice test. Of the 554 nurses who commenced the course, 334 completed the test, and 229 passed it.

Data Collection

Once a participant completed and passed the test, the course tutor invited them by e-mail to respond to a postcourse questionnaire. The questionnaire was sent as a Word file attached to the invitation. Questionnaires were returned by e-mail to the course provider, so were not anonymous.

Instrument

The questionnaire (available, in Italian, from the authors) consisted of 13 questions with closed response choices. It was designed to identify reasons for a trainee's choice to participate in an online rather than a classroom-based course; preferences for the times of day in which courses and tutors might be available, print- or screen-based courses, and courses with or without a tutor; and options for inclusion of collaborative learning activities.

Sample

Of the 229 nurses who successfully completed the course, 152 replied to the questionnaire, a response rate of 66%. The sample consisted of 52 (34%) males, and 100 (66%) females. Most completed the course (including the final test) within nine days of commencing. Almost 50% were aged between 31 and 40, 25% were between 41 and 50, and 10% were over 50.

FINDINGS

Participation Patterns

The majority of the respondents (99, 65%) participated in the course most often from home, while 27% (41) participated most often from their place of work. Most worked on the course in the afternoons or evenings (59, 29% between 2:00 p.m. and 8:00 p.m., and 65, 39% between 8:00 p.m. and midnight). A smaller number participated in the mornings (22, 14% between 7:00 a.m. and 2:00 p.m.) and very few in the late night/early morning (6, 4%).

Perceived Advantage of Online Courses

Almost all participants (90%) reported that a significant advantage of an online course was the convenience of studying at times that suit them (Table 1). Over a third of participants (38%) also believed that completing an online course would save time, and a similar proportion (34%) thought that online courses provided easy access to information.

Preferences

The majority of participants (111, 73%) rated online courses as their preferred method of acquiring the necessary CME credits. Most of the remaining participants (34, 22%) would prefer a mixed mode course, with some classroom sessions as well as online learning. Very few of these online course participants (6, 4%) would prefer to attend courses held entirely in the classroom.

Only a small proportion of participants preferred to follow course contents only online on screen. The majority preferred to print all or part of the material. See Table 2.

There was strong support for the availability of a tutor. Most respondents (128, 85%) were unequivocal in their support, a smaller proportion (22, 14%) were indifferent while very few (2, 1%) saw no need for a tutor at all. Most respondents (119, 78%) preferred to communicate with the tutor by e-mail rather than other means, but nearly 20% (29, 19%) would like more interactive methods (such as chat or videoconference) in addition to e-mail.

Attitudes to Collaborative Learning

Most participants would accept some form of collaboration with other participants in their online courses (Table 3). While just over a third (53, 35%) expressed unconstrained interest in collaboration, a higher proportion (85, 55%) was more circumspect, expressing interest in collaboration "only when necessary." A small proportion (13, 9%) did not want courses to include collaboration under any circumstances.

Those participants who expressed an interest in collaboration were asked which online techniques they would prefer to use for collaboration. Their responses are summarized in Table 4. 10% (15) of the respondents were not familiar with the techniques. Among the remaining 90% of respondents, the most favoured technique was the discussion forum (74, 49%), followed by chat (43, 28%) and virtual classrooms (36, 24%), and finally, videoconference (31, 20%).

While participants were interested in some form of collaboration with other nurses, the respondents were divided over the role of collaboration in activities that required practice of a given technique or method. About a third of respondents (59, 38%) would prefer to complete such activities individually, while a similar number (54, 36%) would prefer to collaborate with other participants. The majority of the remaining respondents (33, 22%) would prefer to interact only with the course tutor.

Participation in the Course Forum

Trainees posted 19 messages to the course forum over the two-month life of the course. Half the messages were questions that were answered by the tutor. Five messages were about administrative issues that were resolved by the tutor. Six messages were from participants expressing their pleasure at successfully completing the course and expressing their appreciation of the way the course was structured and run. Seven messages over 15 days toward the end of the course period discussed the quality of documentation for health professionals working in issue.

DISCUSSION

Before considering the results in detail, we need to consider the sample of course participants who responded to this study. The respondents consisted of two-thirds of those who had successfully completed an online course. Their views probably represent those who successfully negotiated the online course well, but we do not know if their views are representative of those 220 nurses who commenced the course but did not complete it, or of the 105 who completed the course but did not pass it. In discussing the results, we will, therefore, consider what we can learn from participants who successfully completed a course before considering what research still needs to be done.

The strong preference expressed for online courses among those who completed the course indicates that online learning is appropriate for at least some nurses. While there was some demand for classroom-based courses and for courses that involve a blend of online and classroom discussion, such demand may be lower among participants who have been able to successfully complete a course online than others. Initiatives (such as the Italian pilot courses) that promote online courses, therefore have an important role in educating participants about the possibilities of online learning, and should be designed not just to test technology or market, but also in such a way that participants are motivated and supported to complete successfully.

Table 1. Relative advantage of online courses

Potential advantage	*n*	%
Convenience of studying at times that suit me	137	90%
Time saving	58	38%
Ease of access to information	52	34%

Table 2. Preference for print or screen

Preference	*n*	%
Print everything	59	39%
Print only what interests me	49	32%
Save the parts that interest me on my PC	39	26%
Read all on the screen	6	5%

Table 3. Preference for collaboration with other participants

Preference for collaboration	*n*	%
Yes, with pleasure	53	35%
Yes, but only if necessary	85	55%
Not under any circumstances	13	9%
No response	1	1%
Total	152	

Table 4. Preferred technologies for collaboration

Preferred technology	*n*	%
Discussion forum	74	49%
Chat	43	28%
Virtual classroom	36	24%
Videoconference	31	20%
None of these	7	5%
I am not familiar with these techniques	15	10%
Total responses	152	

Although there was little opportunity for collaborative learning in the pilot course described here, 90% of participants declared some interest in collaboration, more than a third of them without qualification. Nonetheless, more than half the participants pointed out that they would participate in collaborative learning activities only if they were "necessary." If collaborative learning is included in online courses, the reason for its inclusion and the value of participation in collaborative learning activities should be clear to the participants. We do not believe that the mere provision of a forum or "café" for discussion is sufficient to motivate the majority of participants to engage in social learning. Rather, collaborative learning activities should be incorporated in the curriculum in such a way that the relationship between the collaborative activities and the learning objectives for the course is clear.

There are also differences in preferences for forms of collaboration. Around half the participants in this study expressed an interest in participating in asynchronous discussion forums, but only a quarter expressed an interest in synchronous methods such as chat, virtual classrooms, or videoconferencing. This preference for asynchronous over synchronous communication is consistent with the participants' appreciation of online learning for the control it gives them over the time when they study. What do we do about the approximately 10% who would prefer never to engage in collaborative activities online? If, even when faced with clear learning objectives for collaborative activities, these people would not participate, the online collaborative activities of other participants would be disrupted.

These differences in preferences for collaborative learning may point to three different market segments for online CME: nurses who would prefer a traditional online course in which material is delivered directly to a learner working alone; nurses who appreciate collaborative opportunities that clearly add to the learning offered by the course, but who prefer those collaborative opportunities to use asynchronous methods that permit them to participate in courses at times that suit them; and nurses who would appreciate courses that use synchronous technologies that provide greater opportunity for collaboration. Even for this last group, we suspect that synchronous technologies would be most useful if they allowed the maximum flexibility in the timing of the student's participation in the course. For example, activities could be designed to permit participants to initiate chat and videoconferencing sessions at times that suit them. Virtual classroom sessions would compel participants to attend class (albeit at a distance) at the time the classes were offered, unless a system (such as Lectopia, http://lectopia.uwa.edu.au/) for recording and replaying the class is made available. Given the preference for asynchronous collaboration, and the range of hours in which participants prefer to study online, courses that include virtual classroom sessions should be designed to include recording and provision of any time access to the virtual classes.

There remains the issue of what type of course would be suitable for those participants who did not enroll in the online course, complete it, or pass it. Here, in the absence of further research, we can only speculate. Evidence from other courses suggests that lack of access to, or familiarity with, course technology can act as a significant barrier to participation (Buckingham, 2003). We can speculate that some potential participants did not enroll and others commenced, but did not continue simply because they were unable to overcome technological barriers (or felt they were unable to overcome them). Online learning might still be appropriate for these participants, but the introduction of online learning should include assistance with and orientation to the technology, not just the course itself. Why did 40% of those who began the course not finish it? Again, the literature offers some suggestions. If it is true that many learners are motivated by learning socially (Levy, 2007), a course that offers no more than a series of screen-based modules with optional tutor assistance and forum participation may not have been sufficiently motivating. The qualified interest in collaborative learning expressed by those who, in any case, completed the course suggests that inclusion of collaborative exercises may assist with participant motivation. Would formalized collaborative learning have improved the pass rate in the course? Again, the general literature on collaborative learning suggests that this would be the case (Rudestam & Schoenholtz-Read, 2002), but further research specifically among nurses undertaking CPE is needed to confirm the role of collaborative learning, both in motivating and improving the learning of participants.

CONCLUSION

This small study confirmed that online distance learning is a feasible, and even preferred, option for some

Italian nurses. Since all health professionals in Italy are required to participate annually in continuing professional education, online distance learning is likely to be an important—and appreciated—way to learn and to meet accreditation requirements. Moreover, many participants value social contact, both with the tutor and with fellow participants. There appear to be three market segments for online CME for nurses: nurses who prefer to study alone, those who would appreciate collaborative activities well-integrated into course design, and those who would prefer courses that include online collaboration of any kind.

How large is the demand for online distance learning in a system where some form of compulsory CPE is required for continued accreditation as a healthcare professional? While we know the preferences of those who already have successfully experienced online learning, we need to confirm that these preferences are shared by nurses who do not have similar experiences. Future research will need to include nurses who have not participated in online learning before we can find out.

REFERENCES

Abdulrasul, S. (2001). *An evaluation of the use of the NHSnet in GP surgeries in terms of user satisfaction.* Unpublished as cited in Klecun & Cornford (2005), Brunel University, London.

Buckingham, S. (2003). Perspective on the experience of the learning community through online discussions. *Journal of Distance Education, 18*(2), 74–91.

Contreras-Castillo, J., Favela, J., Perez-Fragoso, C., & Santamaria-del Angel, E. (2004). Information interactions and their implications for online courses. *Computers & Education, 42*(2), 149–168.

Diekelmann, N., & Mendias, E. P. (2005). Being a supportive presence in online courses: Knowing and connecting with participants through writing. *Journal of Nursing Education, 44*(8), 344–346.

Howatson-Jones, L. (2004). Designing Web-based education for nurses. *Nursing Standard, 19*(11), 41–44.

Howcroft, D., & Mitev, N. (2000). An empirical study of Internet usage and difficulties among medical practice management in the UK. *Internet Research: Electronic Networking Applications and Policy, 10*(2), 170–181.

Kenny, A. (2000). Online learning: Enhancing nurse education? *Journal of Advanced Nursing, 38*(2), 127–135.

Klecun, E., & Cornford, T. (2005). A critical approach to evaluation. *European Journal of Information Systems, 14*(2), 407–421.

Levy, Y. (2007). Comparing dropouts and persistence in e-learning courses. *Computers & Education, 48*(2), 185–204.

Ministero della Salute Italy. (2000, September 14). *Il programma nazionale di Educazione Coninua in Medicina*. Retrieved February 11, 2008, from http://ecm.sanita.it/presentazione/programma.htm

Parker, A. (2003). Identifying predictors of academic persistence in distance education. *USDLA Journal, 17*(1), 55–62.

Platzer, H., Blake, D., & Ashford, D. (2000). An evaluation of process and outcomes from learning through reflective practice groups on a post-registration nursing course. *Journal of Advanced Nursing, 31*(3), 689-695.

Renzi, S., & Klobas, J. E. (2002). *Developing community in online distance learning.* Paper presented at the Xth European Conference on Information Systems. Gdansk, Poland.

Ricketts, C., Price, J., & Chamberlain, S. (2005). Interactive online assessment training for busy practitioners. *Medical Education, 39*(5), 525–526.

Rudestam, K. E., & Schoenholtz-Read, J. (2002). The coming of age of adult education. In K. E. Rudestam, & J. Schoenholtz-Read (Eds.), *Handbook of Online Learning: Innovations in Higher Education and Corporate Training* (pp. 3-28). Thousand Oaks, CA: Sage.

Scheepers, R., Scheepers, H., & Ngwenyama, O. K. (2006). Contextual influences on user satisfaction with mobile computing: Findings from two healthcare organizations. *European Journal of Information Systems, 15*(3), 261–268.

Wiredu, G. O., & Sorenson, C. (2006). The dynamics of control and mobile computing in distributed activities. *European Journal of Information Systems, 15*(307–319).

KEY TERMS

Collaborative Learning: Learning that occurs through the exchange of knowledge among learners.

CME Credits: A system for calculating the amount of CME a healthcare professional has undertaken. In many countries, or for many professional bodies, a minimum number of CME credits must be obtained in a given period to retain accreditation to practice.

Continuing Medical Education (CME): Postqualification professional development, education, and training activities undertaken by doctors, nurses, and other healthcare workers in order to keep their knowledge up to date.

Continuing Professional Education (CPE): Postqualification education undertaken by professionals to ensure they keep up to date with developments in their field.

Online Collaborative Learning: Learning that uses the Internet and Internet-enabled software tools to support social and collaborative learning among students at a distance from one another and from their instructor.

Online Learning Activities: Learning activities in which students interact with resources, or other students, or both, using the capabilities of the Internet or other computer-based communication networks.

Social Learning: Learning through social interaction with other people.

Optimization of Medical Supervision, Management, and Reimbursement of Contemporary Homecare

B. Spyropoulos
Technological Education Institute of Athens, Greece

M. Botsivaly
Technological Education Institute of Athens, Greece

A. Tzavaras
Technological Education Institute of Athens, Greece

K. Koutsourakis
Technological Education Institute of Athens, Greece

INTRODUCTION

The concepts of health, sickness, and illness are subject to the specific sociocultural conditions under which they are considered, and, on the basis of which medical care is provided (Spyropoulos & Papagounos, 1995). The concepts and the methods involved in diagnosis and treatment are subject to the prevalent at the time theoretical model of disease. Hospitals, as a social institution, emerged as a response to particular needs and corresponded to the specific level of the understanding of health and disease. About 2,500 years ago, the temples of Asklepeios, the god of medicine, were probably the first well organized houses of refuge for the sick and training schools for physicians. Hospitals also existed in India under Buddhist auspices as early as the 3rd century BCE. The number of hospitals grew in the first centuries of the Christian era. In the 4th century AD, hospitals were founded in Caesarea and in Rome. Throughout the Byzantine time, the Middle Ages, the Arab and Ottoman dominance periods, the Renaissance, and even later, hospitals were almost entirely run by religious Christian or Islamic groups. During all these centuries, home care remains the main and usually the unique mode of treatment for the majority of the people world-wide. Only during the 18th century, hospitals ceased to be purely philanthropic institutions and they started to assume the character of a social institution where a systematic and theory-infused approach to disease prevailed. From the middle of the 19th century on, the number of hospitals, particularly in Europe and in the USA, increased, principally because of the discovery of anesthesia, aseptic surgical techniques, and, by the end of the century, the introduction of the x-rays. The demand for hospital services expanded further with the spread of prosperity, and with the introduction of various forms of hospitalization insurance, especially in England and in Germany, where the first obligatory and generalized social insurance system is introduced. It is the first time in the human history that home care ceased to be the main way of providing healthcare, and hospital treatment became gradually an important social right.

The modern hospital emerged gradually and successively, during a very long historical development, from a religious philanthropy institution to the contemporary managed care Establishment. The civil structure, the social demands, and the individual performance were always and are still reflected, on the hospital, throughout the centuries. Therefore, the 21st century hospital will provide a radically different professional activity environment, and a quite different professional-patient interaction modus; it will increasingly encourage telemedicine (De Leo, 2002) supported home care because of the increase of mean life expectancy, and the hospital care cost avalanche (Wipf & Langner, 2006; Woolhandler, Campbell, & Himmelstein, 2003). Its mission will be completed by a network of various associated Institutions, providing care rather closer to home care, than to that of the traditional hospital-care (Brazil, Bolton, Ulrichsen, & Knott, 1998).

Adapting medical and managerial decision-making (Spyropoulos, 2006a) in the modern home care environment is a cardinal prerequisite, in order to ensure, first, an economically sustainable development of the aging population healthcare (Scarcelli, 2001); second, the rehabilitation services required for impaired persons; and finally, the psychosomatic support necessary in the developed countries, during the next decades. Thus, a strategic question emerges that is how home care will be medically supervised and financially reimbursed. The present study attempts to describe the present situation and the contemporary technological trends in home care; more specific, it is focused on a system developed by our team that intends first, to enable the optimal documentation of the provided home care, and second, to facilitate the acquisition of all relevant financial data, leading to a fair remuneration of the services offered.

BACKGROUND

Contemporary home care is evolving on the foundation of a variety of applications in healthcare support services, and we argue that a qualitatively new "mobile" home care is presently emerging out of the combined employment of, first, the modern wireless mobile telephony networks and equipment, second, the contemporary digital entertainment electronics, and third, the commercially available computer hardware and software. This new mobile home care allows us for to be optimistic about the reduction of patients' unnecessary hospitalization in the near future, as well as, the dramatic reduction of the home care costs. It is essential to summarize the most important emerging innovative aspects of modern home care before describing in detail the developed therefore Management System.

Adapting medical decision-making and treatment in the home care environment begins with the effort to carry out periodically preventive examinations, as a kind of ambulant and emergency home care. The miniaturization of equipment, and the falling prices trend that could be developed by the opening of Biomedical Technology to the immense population of potential consumers, if combined with a "smart" home environment, then it may result in periodical, individually adapted, preventive medical examinations, and, consequently, a decreasing number of patients in hospitals and other costly traditional ambulatory services. Diagnosis and Treatment of an emergency patient at home is inevitable, in order to reduce morbidity, following acute pathological reasons or an accident, and it is paramount that the patient is first evaluated and stabilized on site. In such cases, there are two complementary ways to intervene: First, the continuous monitoring of high-risk groups suffering from chronic cardiovascular, respiratory, and other diseases; and second, equipping and training the "first responder" with appropriate hardware and standardized guidelines (Spyropoulos, 2006b). These two main options allow for an improvement of the efficiency of ambulant emergency services, especially if supported by telematics.

Concerning home-based in vitro diagnostics, "Dry Clinical Chemistry" offers today a variety of products, covering the whole range of important parameters, like metabolites, enzymes, electrolytes, and so on. Although diabetes every-day strip-based monitoring (Lewis, 2001) is a routine for the last two decades, a little has been done to establish a "home- based" in vitro diagnostics laboratory that would be very useful, especially in facilities offering housing and some kind of medical care for elderly people (Gill, 2002). In vitro testing, combined with biosignals monitoring at home, provide for a safe and pleasant living environment for sensible populations. Concerning home imaging, although several mobile systems have been developed, most of them, with the exception of ultrasonic equipment, will remain for the near future the "status symbols" of the hospitals. However, external patient-imaging, together with high-quality biological sound processing and transmission capabilities, are already available.

Although surgery will remain more or less the "monopoly" of the hospital, the postoperative hospital stay will be further dramatically shortened and replaced by home care in a familiar and pleasant room, instead of the impersonal, sterile ward. This becomes feasible, first, because of the recently developed minimally invasive surgical techniques, and second, as a result of decentralizing postoperative care to general practitioners and nurses, and finally, due to the emerging option of, the acquisition at home of vital-signs, postoperative wound images, and so on, and their transmission to the specialized surgical center, if necessary.

Although "home-ICU" sounds like a joke, there are always circumstances that may lead to interim intensive care settings, for instance on an uneasily accessible site after an accident. Virtually, within a telematics supported home care environment, the patient's monitoring, until safe transportation becomes possible, is not a major problem. Even Cardio Pneumonary Resuscitation (CPR) and defibrillation are feasible options. However, the access to specialized human and material resources will keep the ICU, as a hospital dedicated department, for the years to come. Furthermore, in cases like Chronic Obstructive Pulmonary Disease (COPD), Chronic Asthma, and other such diseases that cause air-flow limitation, leading to inability to carry out daily activities, an appropriate home care environment may contribute, first, to home-treatment patient-customized programs, aiming to improve the subject's performance and, thereby, the quality of life; second, to the evaluation of preceding pharmacological or surgical treatment; and third, to the definition of optimal selection criteria for home-based assisted ventilation (Simonds, 2003).

Several entertainment and medical electronics features already developed or under development, can easily be adapted to serve persons with limited mobility, poststroke patients, patients suffering from certain neurological and/or mental disorders. Such features include, among others, first, Web camera-supported presentations on interactive high-definition monitors, creating the feeling of being together, when physically separated from relatives and friends; second, every-day life equipment remotely controlled, third, electrical stimulators, developed by pacemaker manufacturers that improve poststroke incontinence control, reduce tremor on Parkinson's disease patients, and so on. Finally, interactive Internet and TV could contribute, first, to the reduction of the enormous rehabilitation costs, by providing online instructions to home-based post-traumatic or poststroke rehabilitation exercises; second, to the restriction of the agoraphobic symptoms, and of the self-respect-less behavior of depressive patients; and finally to the continuous home-treatment and/or to the special education of patients, especially young ones. However, the presently major difficulty is not the hardware and telecommunication infrastructure, but the development of appropriate content and user interfaces, serving reliably the aims mentioned.

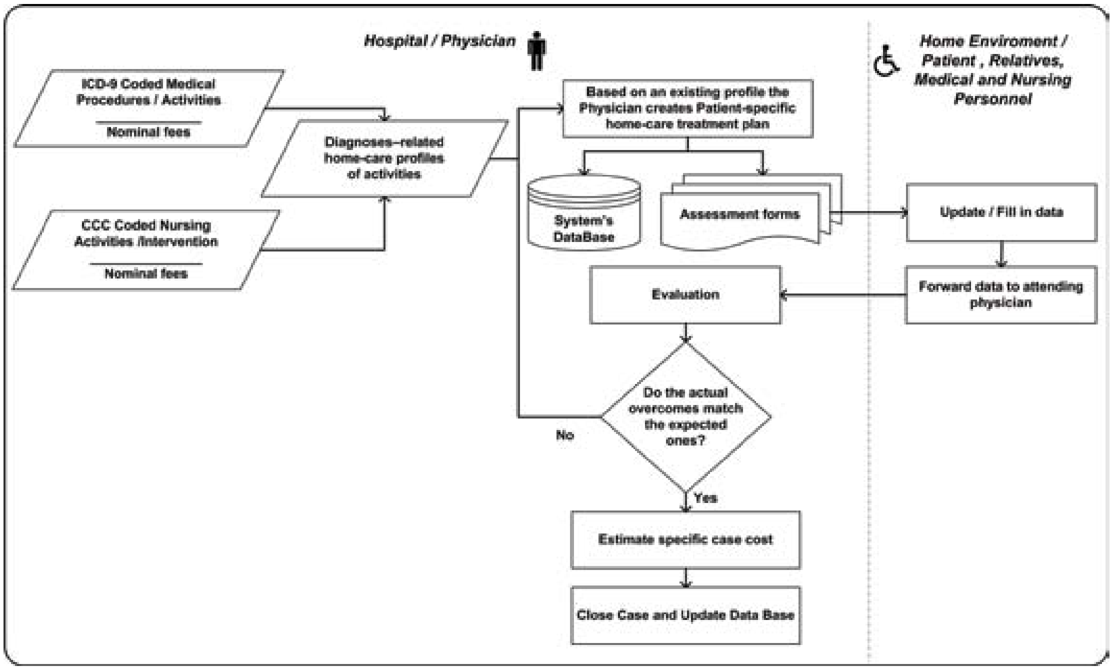
MEDICAL SUPERVISION, MANAGEMENT, AND REIMBURSEMENT OF CONTEMPORARY HOME CARE

The developed system attempts to relate home care to the Diagnosis Related Groups (DRGs) based Reimbursement (Botsivaly, 2006) by introducing an innovative adapted Continuity of Care Record (CCR). The proposed solution consists of a typical CCR-system, equipped with an extension that, first, acquires and processes an individually designed structured set of data to be employed during the post-discharge home care period, and second, allows for the estimation of the homecare cost vs. the DRGs code issued at discharge from the hospital, supporting, thus, a more accurate DRGs-related home care remuneration. Medical records are indispensable in home care, and are also used in a variety of ways, serving a multiplicity of purposes (Papagounos & Spyropoulos, 1999). A major problem of the contemporary health services, including home care, is the scarcity of resources. Patients' records provide information concerning the expenses incurred and the resources allocated for the treatment of the individual patient within the healthcare system. Home care resources consumption must also be monitored, and we argue that the most efficient and time saving way is through the employment of an appropriately adapted Continuity of Care Record (CCR).

The Continuity of Care Record (CCR) is a standard for exchanging basic patient data between two care providers, in order to enable the second one to access relevant patient information. It is intended to foster and improve continuity of patient care, to reduce medical errors, and to assure at least a minimum level of health information transportability when a patient is referred, transferred, or otherwise goes to another provider. The standard was proposed and developed by the E31 Committee on Healthcare Informatics of ASTM, an American National Standards Institute (ANSI) standard development organization (ASTM E2369-05, 2005).

The CCR is being developed to cover the need to organize and make transportable a set of basic patient information consisting of the most relevant and recently facts about a patient's condition. These include, first, patient and provider information; second, insurance information; third, patient's health status such as, allergies, medications, vital signs, diagnoses, recent

Figure 1. Details of system's flow chart



procedures; fourth, recent care provided, as well as recommendations for future care (that is, a care plan); and finally, the reason for referral or transfer. The structure of the CCR perfectly fits to the needs of homecare, both, the administrative and the medical ones, because it is designed to be technology- and vendor-neutral for maximum applicability. It is being developed on the extensible markup language (XML) platform in order to offer multiple options for its presentation, modification, and transmission.

Through XML, the CCR can be prepared, transmitted, and viewed in a browser, in an Health Level Seven (HL7) Clinical Document Architecture (CDA) compliant document, in a secure e-mail, or in any XML-enabled word processing document. This makes it possible for recipients to access and view the information in the manner that they prefer (electronic or paper), to extract the data as required, and even to store them on a portable storage device for use as a personal health record. Because the CCR will be a simple XML-document, different Electronic Health Record **(**EHR) systems will be able to, both import and export all relevant data to and from the CCR document, and enable automated transmission with minimal workflow disruption for individual caregivers. Thus, the CCR will increase interoperability between different EHR systems, with minimal or no custom programming. Finally, in countries with minimal EHR and healthcare networks infrastructure, the CCR is the only advisable interim operational mode.

We have designed and developed a first version of a CCR-extension for comprehensive payer-specific information. The developed model consists of various related databases, and allows for every user, first, to select one of the available DRG-codes; second, to create a typical CCR that contains the appropriate demographic and administrative data, as well as, the relevant clinical information; third, to set up for each DRG code a custom-made typical profile of home care activities; and fourth, to attach them either a nominal fee, proposed by the system, which is based upon realistic, empirically collected data, or to independently price them. The developed system employs the Australian Refined DRGs (AR-DRGs).

Starting point is the loading of a number of well established medical classification systems, like the International Classification of Diseases Version 9 (ICD9), the Australian Refined DRGs (AR-DRGs) and the Nursing Interventions Taxonomy of the Clinical Care Classification (CCC) databases (Saba, 2002). The AR-DRGs database figures out about 660 diagnosis-related groups, classified within 25 Major Diagnosis Categories (MDC), 30 Specialties, and three major groups (medical, surgical, and other). From this root databases, every user is able to individually assign an appropriate set of care activities to specific diagnosis codes that are coded according to AR-DRGs. These activity sets consist of diagnostic, monitoring and treatment activities that can be actually performed in a home environment, together with an appropriate nursing—activity treatment plan. These profiles of care activities are custom-made and every user (i.e., every physician responsible for discharging a patient from the hospital) is actually able to set up his own profiles. During the formation of these profiles the user can attach to each activity a set of nominal fees. This later is estimated by another already developed software tool (Botsivaly, 2003), and allows for a rational approximation of the effective mean cost for several elementary medical activities, over different medical specialties. Thus, the developed system ignites, when relevant, the corresponding revision of an implicitly associated latent financial record that allows for an approximation of the individual case-cost.

Once the modified database for specific homecare activities has been created, the program constructs a custom-made Structured Query Language (SQL) file, appropriate for Medical and administrative use. The user can interactively modify any description on this database, and it is possible to add any information for specific instances, in order to adapt the home care profile to emerging new needs. This interactively modified database is further used to create each time a typical CCR that contains all the necessary patient information. The appropriate patient and administrative data, as well as, the relevant medical information, are acquired, either by using an already installed EHR system or even manually, if no EHR is available. Then, the individual CCR is completed, by selecting the needed home care activities, which are automatically inserted in the CCR-section of Care-Plan. Simultaneously the CCR-system latently combines each of them to the predefined "cost."

However, the system, apart from producing, electronically or in paper-format, the CCR, it also produces a number of additional forms, including advisory and informational notes for the patient himself, or for his relatives, and diagrams of physiologic measurements, such as glucose, blood pressure, and so on that the patient should monitor. The system also provides for forms that will be filled by the nursing personnel during the care-visits, in order to document their activities.

Figure 2. Homecare activities selection for a specific patient

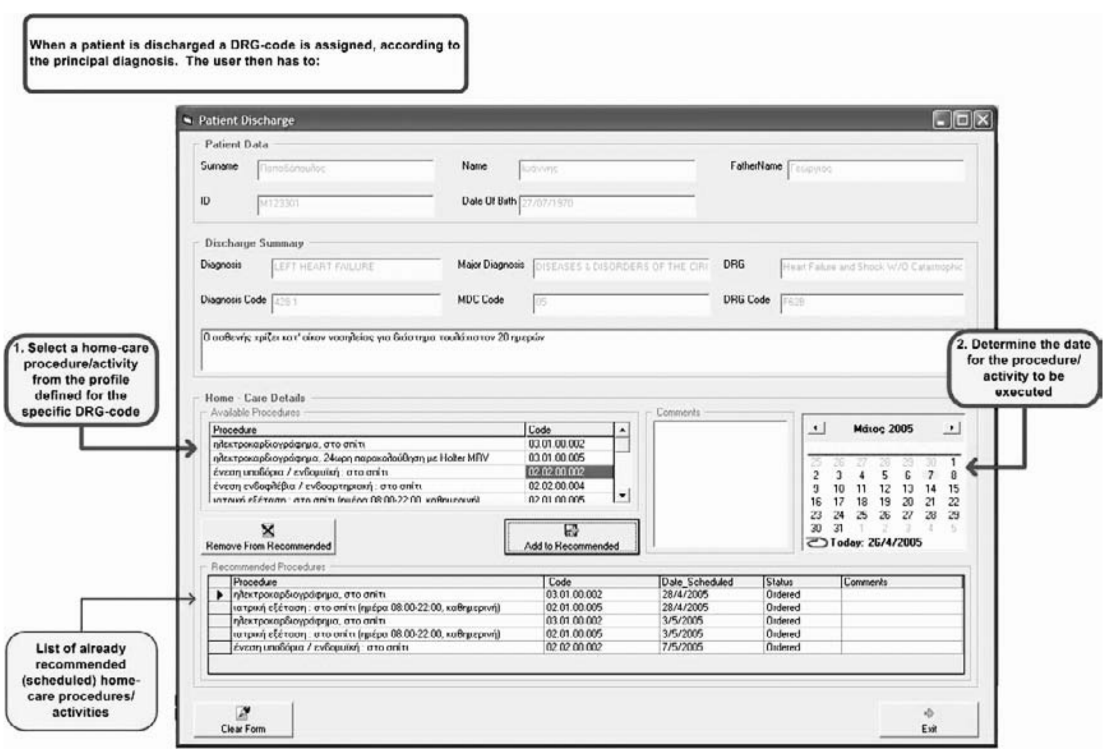
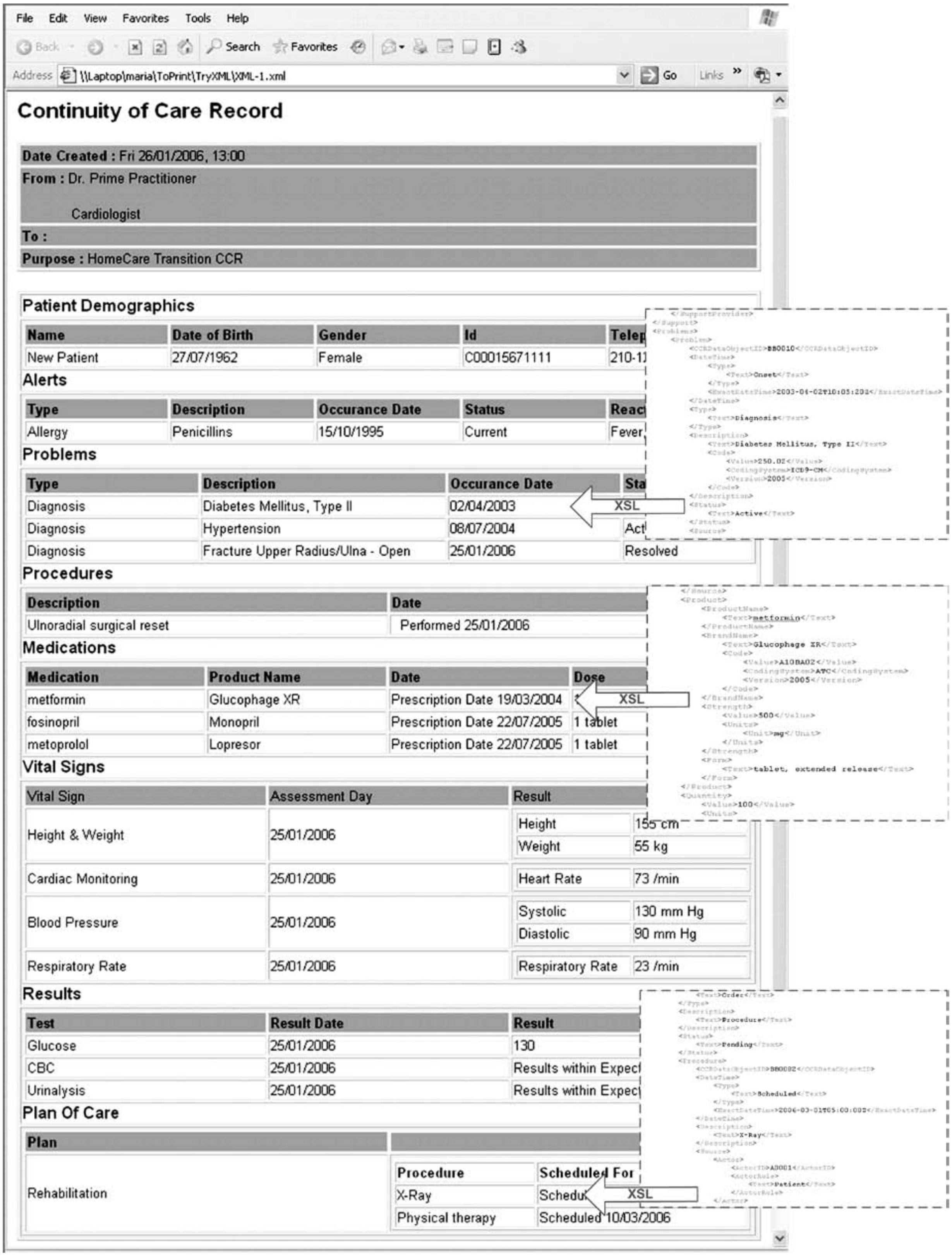


Figure 3. XML to HTML transformation of the produced CCR



Continuity of Care Record

Date Created : Fri 26/01/2006, 13:00
From : Dr. Prime Practitioner
Cardiologist
To :
Purpose : HomeCare Transition CCR

Patient Demographics

Name	Date of Birth	Gender	Id	Telep
New Patient	27/07/1962	Female	C00015671111	210-1

Alerts

Type	Description	Occurance Date	Status	Reac
Allergy	Penicillins	15/10/1995	Current	Fever

Problems

Type	Description	Occurance Date	Sta
Diagnosis	Diabetes Mellitus, Type II	02/04/2003	
Diagnosis	Hypertension	08/07/2004	Act
Diagnosis	Fracture Upper Radius/Ulna - Open	25/01/2006	Resolved

Procedures

Description	Date
Ulnoradial surgical reset	Performed 25/01/2006

Medications

Medication	Product Name	Date	Dose
metformin	Glucophage XR	Prescription Date 19/03/2004	
fosinopril	Monopril	Prescription Date 22/07/2005	1 tablet
metoprolol	Lopresor	Prescription Date 22/07/2005	1 tablet

Vital Signs

Vital Sign	Assessment Day	Result	
Height & Weight	25/01/2006	Height	155 cm
		Weight	55 kg
Cardiac Monitoring	25/01/2006	Heart Rate	73 /min
Blood Pressure	25/01/2006	Systolic	130 mm Hg
		Diastolic	90 mm Hg
Respiratory Rate	25/01/2006	Respiratory Rate	23 /min

Results

Test	Result Date	Result
Glucose	25/01/2006	130
CBC	25/01/2006	Results within Expec
Urinalysis	25/01/2006	Results within Expec

Plan Of Care

Plan	Procedure	Scheduled For
Rehabilitation	X-Ray	Schedu
	Physical therapy	Scheduled 10/03/2006

The filled forms, both the ones regarding the nursing activities and interventions, and the ones regarding the monitoring of physiological parameters, are returned to the responsible physician who evaluates them and, depending on his evaluation, can modify the care-plan of the specific patient, in any suitable way. The structure and data of the produced CCR are complying with the ASTM E2369-05 Specification for Continuity of Care Record, while XML is used for the representation of the data. The XML representation is made according to the W3C-XML schema proposed by ASTM. The CCR that is produced by the system, is currently automatically transformed to HTML format, using the Extensive Stylesheet Language (XSL), in order to be viewable and printable.

FUTURE TRENDS

The previously described employment of some kind of records, as the main instrument for the management of home care, assigns them an additional dimension, which is related to the nature and function of technology. The development of patients' records exhibits the state-of-the-art Biomedical Technology, the dominant—at the time—medical knowledge and practice, and it reflects the existing value system of a given society. The assimilation of new technologies is a complex phenomenon, which is affected by the wider sociocultural, epistemic, and axiological parameters at work, within the specific social context in question.

In other words, the Technologies employed presently in the construction of medical records, were developed to satisfy concrete social needs. The availability of these technologies permitted new questions to be asked and new needs to arise, and this procedure will determine the future trends of the structure, the content, and the technological substratum of the medical records of the future.

CONCLUSION

The employment of this method enables the formation of a CCR, that is intrinsically, however latent, related to the costs caused and the expected reimbursement. Thus, the updating of the patients' relevant medical data, ignites, when relevant, the corresponding revision of an implicitly associated financial record, that allows for, both, first, a good approximation of the individual "case" cost, and second, the follow-up of cost evolution of a hospital-based or freestanding home care group. Thus, the developed system introduces a nondisturbing method for the home care personnel, to combine the collection of the necessary CCR information, and the creation of the associated documentation, with a latent simultaneous acquisition of important financial data. The correlation of these data to the DRG-coded principal diagnosis, allows for the estimation of the home care cost vs. the DRGs code issued at discharge from the hospital.

In view of the fact that there is a clearly higher discharge to home care incidence, for specific DRG-coded principal diagnoses, the developed system is now being tested with an EHR system that has been developed by our team under virtual home care conditions, and is being fed with financial data related to home care activities for some of the most frequent cases. The implementation indicates, so far, that the system, whether interfaced to an EHR or not, is stable enough for practical use, and it actually provides a simple, effective, and expandable tool for the formation of both a CCR and a home care plan, offering at the same time a good approximation of the individual case cost and a flexible HTML-format for data representation. The system allows for the improved approximation of the home care cost vs. the DRGs code issued at discharge from the Hospital supporting, thus, a new more accurate DRGs-related home care remuneration modus. Finally, the system constitutes an effective educational tool, supporting personnel training in operational cost management.

REFERENCES

ASTM E2369-05. (2005). *Standard specification for continuity of care record*. Retrieved February 5, 2008, from http://www.astm.org

Botsivaly, M., et al. (2006). Software development for the estimation of the mean DRGs related treatment cost. *Studies in Health Technology and Informatics, 124*, 307–312.

Brazil, K., Bolton, C., Ulrichsen, D., & Knott, C. (1998). Substituting homecare for hospitalization: The role of a quick response service for the elderly. *Journal of Community Health, 23*(1), 29–43.

De Leo, G., et al (2002). WEB-WAP based telecare. *JAMIA Symposium Supplement 2002*, 200–205.

Gill, T., et al. (2002). A program to prevent functional decline in physically frail elderly persons who live at home. *The New England Journal of Medicine, 347*(14), 1068–1074.

Lewis, C. (2001). Home diagnostic tests: The ultimate house call? *U.S. Food and Drug Administration Consumer Magazine, 35*(6). Retrieved February 5, 2008, from http://www.fda.gov/fdac/601_toc.html

Papagounos, G., & Spyropoulos, B. (1999). The multifarious function of medical records: Ethical issues. *Methods of Information in Medicine, 38*(4), 317–320.

Saba, V.K. (2002). Nursing classifications: Home healthcare classification system (HHCC): An overview. *Online Journal of Issues in Nursing*. Retrieved February 5, 2008, from http://nursingworld.org/ojin/tpc7/tpc7_7.htm

Scarcelli, A.J. (2001). Cost-effective approaches to corporate compliance for hospice and home health providers. *Home Healthcare Management & Practice, 13*(4), 176–183.

Simonds, A.K. (2003). Home ventilation. *European Respiratory Journal, 22*(47), 38–46.

Spyropoulos, B., et al. (2006a). Application tools supporting decision-making in the modern hospital. *International Journal of Scientific Research, 16*, 129–133.

Spyropoulos, B., et al. (2006b). Development of a system supporting patient supervision and treatment in contemporary homecare: Status report. *Studies in Health Technology and Informatics, 124*, 91–96.

Spyropoulos, B., & Papagounos, G. (1995). A theoretical approach to artificial intelligence systems in medicine. *Artificial Intelligence in Medicine, 7*(5), 455–465.

Wipf, K., & Langner, B. (2006). Policy approaches to chronic disease management. *Home Healthcare Management & Practice, 18*(6), 452–462.

Woolhandler, S., Campbell, T., & Himmelstein, U.D. (2003). Costs of healthcare administration in the united states and canada. *The New England Journal of Medicine, 349*(8), 768–775.

KEY TERMS

American Society for Testing and Materials International (ASTM): An international voluntary standards organization established 1898 in the United States by Charles Benjamin Dudley. ASTM develops technical standards for materials, products, systems, and services, and maintains presently more than 12,000 standards that have been incorporated into or are referred to by many federal regulations.

Clinical Care Classification (CCC): A categorization system that consists of interrelated terminologies—the CCC of Nursing Diagnoses and Outcomes and the CCC of Nursing Interventions and Actions—with both classified by 21 Care Components, that classify, and track care based on the six steps of the Nursing Process Standards of Care recommended (1998) by the American Nurses Association (ANA): assessment, diagnosis, outcome identification, planning, implementation and evaluation.

Clinical Document Architecture (CDA): An XML-based markup standard intended to specify the encoding, structure, and semantics of clinical documents for exchange. The CDA tries to ensure that the content will be human-readable and hence is required to content narrative text, yet still contain structure, and most importantly, allows for the use of codes to represent concepts.

Continuity of Care Record (CCR): A health Record standard (ASTM E2369-05) specification that constitutes a patient health summary and contains the most relevant and timely core health information to be sent, usually in electronic form, from one care giver to another. It contains various sections such as patient demographics, insurance information, diagnosis and problem list, medications, allergies and care plan, representing a "snapshot" of a patient's health data that can be useful or possibly lifesaving, if available at the time of clinical encounter. The CCR standard is expressed in the standard data interchange language known as XML and can potentially be created, read and interpreted by any EHR and EMR software applications. A CCR can also be exported in other formats, such as pdf, doc, and so on.

Diagnosis Related Groups (DRGs): A system that classifies hospital cases into one of approximately 500–600 groups, which relate types of patients treated, to the hospital resources they consumed. It was first developed by Robert Barclay Fetter and John Devereaux Thompson at Yale University, in the slate 1970s. The DRGs are assigned by a "grouper" program based on ICD diagnoses, procedures, age, sex, and the presence of complications or comorbidities.

Electronic Health Record (EHR): A personal medical record in digital format, typically accessed on a computer or over a network. An EHR almost includes information relating to the current and historical health, medical conditions, and medical tests of its subject. In addition, EHRs may contain data about medical referrals, medical treatments, medications and their application, demographic information, and other nonclinical administrative information.

Extensible Markup Language (XML): A general-purpose markup language designed to be reasonably human-legible, and therefore, abruptness was not considered essential in its structure. Its primary purpose is to facilitate the sharing of data across different information systems, particularly connected through the Internet, and to allow for diverse software to understand information formatted in this language.

Health Level Seven Inc. (HL7): HL7 is a not-for-profit organization, established 1987, involved in the development of international healthcare standards, accredited by the American National Standards Institute (ANSI). HL7 is currently the selected standard for the interfacing of clinical data for most hospital information systems worldwide.

Homecare: Healthcare provided in the patient's home by healthcare professionals or by family and friends. Homecare aims to enable people to remain at home rather than use institutional-based nursing care. Care workers visit patients in their own home to help them with daily tasks and supervision of treatment. Homecare is generally paid for by private health insurance, by public payers, or by the family's or patients' own resources.

International Statistical Classification of Diseases and Related Health Problems (ICD): The ICD is published by the World Health Organization (WHO), and provides codes to classify diseases and a wide variety of signs, symptoms, abnormal findings, complaints, social circumstances and external causes of injury or disease. Every health condition can be assigned to a unique category and given a code, up to six characters long. Such categories can include a set of similar diseases. The ICD is used world-wide for morbidity and mortality statistics, reimbursement systems and automated decision support in medicine. This system is designed to promote international comparability in the collection, processing, classification, and presentation of these statistics. The ICD is the core classification of the WHO; it is revised periodically, and is currently in its tenth edition (ICD-10).

Medical Classification: The method of transforming descriptions of medical diagnoses and procedures into universal Medical code numbers. These diagnosis and procedure codifications are used by health insurance companies for reimbursement purposes, they support statistical analysis of diseases and therapeutic actions, they are employed in knowledge-bases and medical decision support, and in a variety of other uses.

PACS Contribution to Hospital Strategy via Improved Workflow

Reeva Lederman
University of Melbourne, Australia

Rogier van de Wetering
University of Melbourne, Australia

Lucy Firth
University of Melbourne, Australia

INTRODUCTION

Since Strassman's (1990) exposition of the "productivity paradox," there has been increased attention paid to justification and evaluation of investments in information technologies (IT) and information systems (IS). Observed outcomes from practice have called for theoretical explanations in order to construct a generalized view of IS. In the hospital environment, we see that there is a pressing need for frameworks adequate to the tasks of evaluating increasingly expensive implementations. Because hospitals have multiple stakeholders and multiple strategic goals, in order to be adequate, a framework needs to address these multiplicities. One promising framework is Kaplan and Norton's (1992), the Balanced Score Card (BSC).

BSC has been used to evaluate hospital performance and the implementations of IT/IS in hospital and medical contexts by several authors (Curtright, Stolp-Smith & Edell, 2000; Gordon & Geiger, 2001). However, Picture Archiving and Communication Systems (PACS), which is one of the most important system implementations in the past few years in the area of e-health and the transmission of health data, have not been evaluated using BSC. Rather, most evaluations of PACS have tended to focus on single issues like clinical communication, quality improvement, image availability, speed of service, and workflow simplification and automation, and the associated gains that are important to the patient's overall journey (Peer, Peer, Walcher, Pohl & Jaschke, 1999). The focus on single issues within PACS implementations has made it difficult to gain a clear understanding of the overall workflow impacts of a PACS implementation. Moreover, the literature does not relate well to the intangible value created for the patient by the patient care benefits of PACS, which is an important component of the hospital's strategy overall and as it relates to IS strategy.

Therefore, this chapter investigates the adequacy of BSC for a holistic evaluation of the workflow impacts of a PACS implementation. It asks whether a theoretical model such as BSC adequately captures the reality of how such technology is used. The approach taken is radical in that it is built on a consideration of the fundamentals of hospital strategy. The BSC is then modified to incorporate qualitative themes rather than performance measures to reflect the fundamentally qualitative nature of the clinical values of hospital strategy. In so doing, this chapter develops a framework that is relevant to a hospital's not-for-profit and clinical strategies.

BACKGROUND

The BSC is a set of measures that provides managers with a comprehensive framework that translates a company's strategic objectives into a coherent set of performance measures. As originally developed by Kaplan and Norton (1992), the BSC includes performance measures from the following four perspectives, supplementing the financial perspective with those of the internal business process, the customer, and learning and growth within the organization:

- **Financial perspective.** "How should we appear to our shareholder?" Performance measures include operating income and return-on-investment.
- **Internal business process perspective.** "At which business processes must we excel?" Performance measures include rework rates, cycle times, and process costs.

- **Customer perspective.** "How should we appear to our customers?" Performance measures include customer satisfaction and retention.
- **Learning and growth perspective.** "How will we sustain our ability to change and improve?" Performance measures include employee skills, retention, and satisfaction.

The comprehensive view drawn from these four performance measures can then be presented as a single management report that reflects many of the elements of a company's competitive agenda: becoming customer oriented, shortening response time, improving quality, emphasizing teamwork, reducing new product launch times, and managing for the long term (Kaplan & Norton, 1992). Not only does the BSC provide a measurement framework that improves alignment of actions to the strategic goals of an organization, but it also provides a platform for identifying priorities (Mooraj, Oyon & Hostettler 1999). These priorities can then be used to guide management in the achievement of objectives.

However, the BSC is not a template that can be applied to business in general or even industrywide. Rather, it is intended that different market situations, product strategies, and competitive environments employ different scorecards, differing in terms of performance measures. Each organization's unique reason for an IT/IS implementation and therefore different perspectives on measuring success is reflected in the use of a BSC that includes appropriate performance measures. BSCs are particularly appropriate for organizations in industries such as health care, where there is a more diverse set of performance measures than in the business and academic sectors (Voelker, Rakish & French, 2001). Therefore, the BSC's design flexibility makes it applicable to the evaluation of a broad range of organizations and implementations, and suitable to evaluation within the health sector. A range of perspectives has been used to generate performance measures used in BSC applications within the health sector, such as patient satisfaction, clinical outcomes, functional health status, and cost to evaluate outsourcing (Schriefer, Urden & Rogers, 1997) . We argue that these adaptations of the BSC to health care are successful because the modifications are in line with the organizational strategies of the health sector; consequently, there is a value in a BSC that is specifically targeted toward PACS.

PERFORMANCE MEASURES THAT FIT HOSPITAL STRATEGY

"Strategy" in the corporate sense, popularized by Porter (1996), is "the creation of a unique and valuable position, involving a different set of activities … different from rivals." Following from Porter's earlier work, three fundamental strategies for competitive advantage are identified: low cost, product differentiation, and niche market (McFarlan, McKenney & Pyburn 1983). Willcocks, Petherbridge, and Olson (2001) expand these to six strategic uses of IT: breakthrough unit costs for customers, service-based differentiation, micromarketing management, shorter time to market, transfer of experience, and new level of partnership. The idea of strategy as a way of positioning the organization so as to attract customers and compete with rivals is central to these approaches.

However, the relevance of corporate strategy to hospitals that have a commitment to clinical excellence and a commitment to public responsibility has been questioned (Liedtka, 1992). Where clinical and not-for-profit considerations are fundamental to organizational strategy, Liedtka (1992) suggests that Andrews' concept of strategy is more relevant: "A pattern of decision in a company that determines and reveals its objectives, purposes, or goals, produces the principal policies and plans for achieving those goals … and the nature of the economic and non-economic contribution it intends to make to its shareholders, employees, customers, and communities" (Andrews, 1987, p. 56).

Given this richer concept of strategy, Andrews (1987) argues that there are four elements to be considered together to determine strategy:

1. What the market wants in terms of industry opportunities and threats—what might we do?
2. The organization's competence—what can we do?
3. The aspirations and values of executives in charge of the organization—what do we want to do?
4. The organization's obligation to society—what should we do?

Liedtka (1992) argues that in the health context, it is clinical as well as executive preferences that must be considered. Liedtka (1992) summarizes these elements emphasizing the fact that the elements of strategy may be classified as either market-driven or nonmarket-driven, and as internal or external (Figure 1). According to Liedtka (1992), the incompatibility of Porter (1996) and his followers' concepts of strategy for hospitals lie in Porter's exclusive focus on market-driven elements.

Clinical excellence and service to the community are key factors for public hospitals (Firth & Francis, 2004). These factors are important determinants of clinicians' acceptance, use, and "ownership" of IT/IS in hospitals. Clinicians' appropriate technologies to the extent that they support and enhance workflow in ways that improve the patient's clinical journey. Attempts to impose IT/IS that do not result in such positive workflow outcomes for patients lead to poor morale and resistance.

Consequently, BSC evaluations of IT/IS implementations in hospitals need to incorporate a wide range of nonmarket performance measures because they reflect the nonmarket elements of the hospital's strategic perspectives. Thus, a model for the evaluation of health systems based on Andrews' (1987) approach seems more appropriate.

In order to develop an appropriate set of performance measures for the BSC in the PACS environment and to understand those measures in a practical context, a study was undertaken of a PACS implementation in a large public hospital in Australia (identity withheld) offering the full suite of medical inpatient and outpatient services.

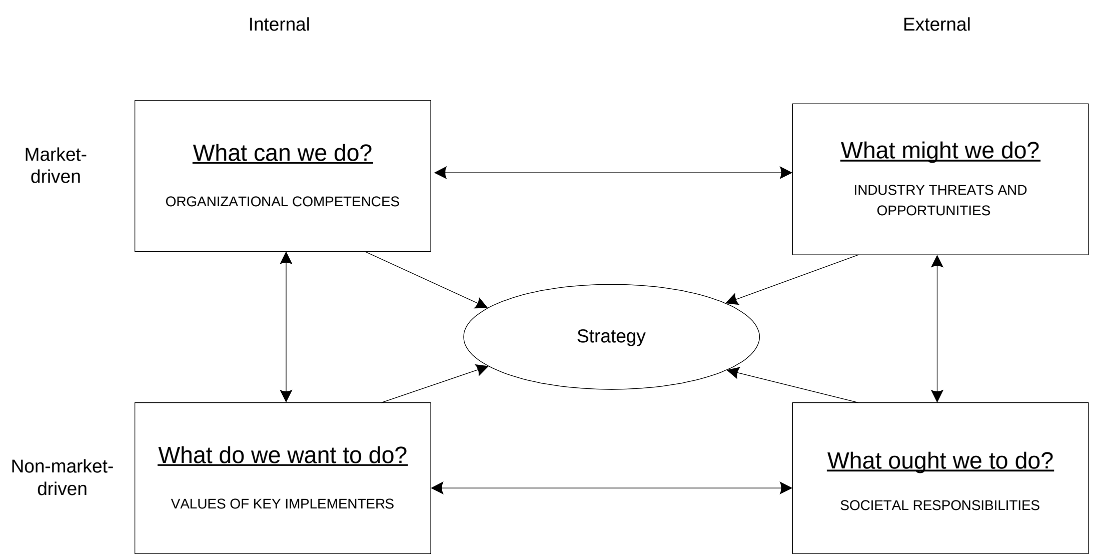
The study involved in-depth, semistructured interviews with users of PACS in several locations throughout the hospital covering three interrelated issues:

1. The goals and strategies of the hospital
2. The importance of workflow to those goals and strategies
3. How well they believed the PACS implementation helped to improve the workflow and subsequently to achieve those goals and strategies

Four measurement perspectives were found. The first two are directly analogous to two of Kaplan and Norton's (1992) original perspectives (internal business processes and customers, respectively). The latter two are more radical modifications to reflect the nonmarket context of the hospital environment in which PACS is implemented:

- A clinical business process perspective that captures the major clinical impacts on workflow has been included, because the outcome—speed of service—is of central importance to workflow analyses. The workflow implications of PACS impacting clinical business processes not only relate to automation but also to the productivity of radiologist, technologist, and clerical staff; report-turnaround times; and changes in communication between radiologist and clinicians. This perspective is consistent with the original

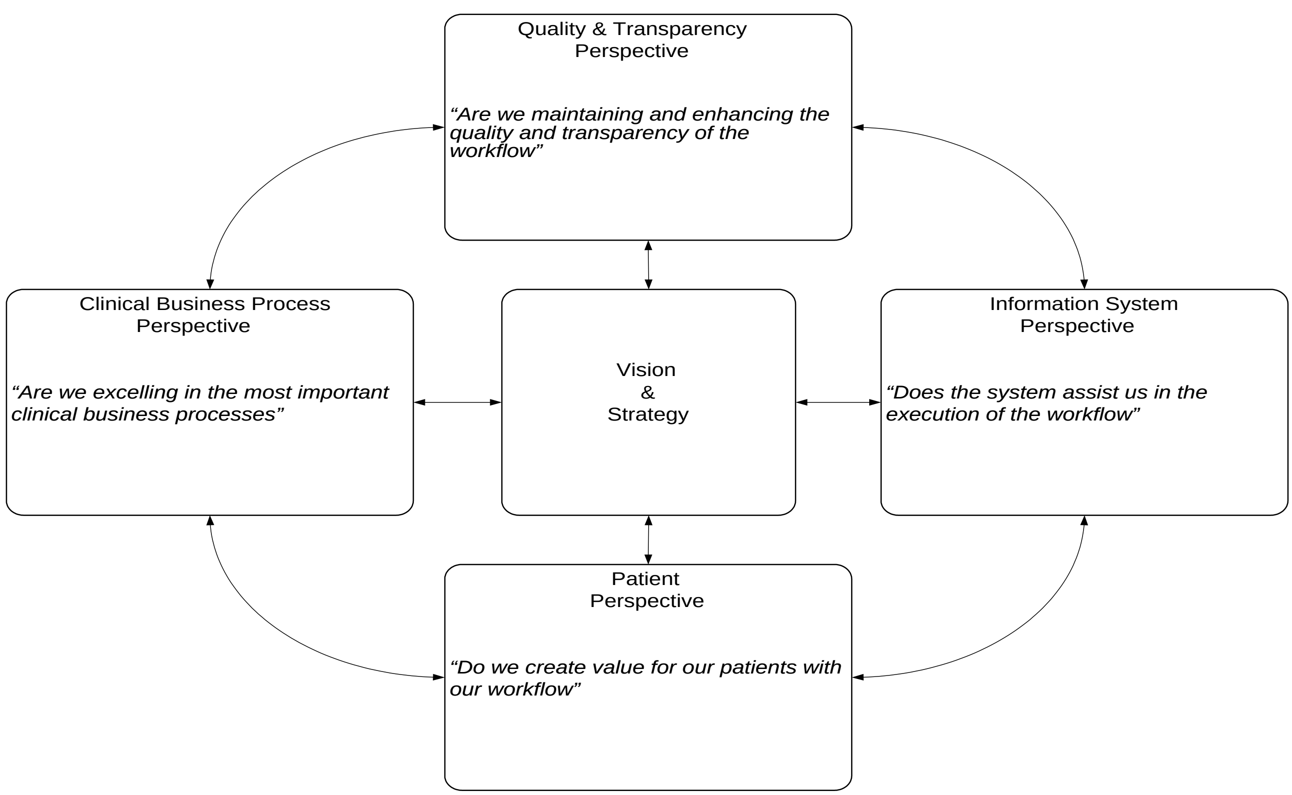
Figure 1. Andrew's model of the strategy formulation process (Liedtka, 1992)



BSC perspective: "At which business processes must we excel?"

- A patient perspective that captures the impact of PACS on the production of value for the patients along the workflow is included, because queues and bottlenecks are symptoms of poor patient throughput. The workflow implications that are directly patient-related include reduction in delays and an increase in timely care, as well as throughput time of patients (Andriole, 2002). This perspective is consistent with the original BSC perspective: "How should we appear to our customers?"
- A quality and transparency perspective that captures the level of monitoring and status checking possible in the system is included, because the literature suggests that improved workflow quality and transparency give higher diagnostic value. This radical departure from Kaplan and Norton's (1992) original four perspectives reflects the fact that hospitals not only need rapid workflow, but they also must have quality checks that imply a quality and transparency of workflow. It reflects an emphasis on Andrews' (1987) perspective of "what do we want to do" and "what should we do?" when hospitals are focused on workflow as a way of enhancing patient care. This perspective asks, "Are we maintaining and enhancing the quality and transparency of our workflow?"
- An information systems perspective that captures the impact that PACS has on the enabling power of the information system is included. This is because the literature suggests that PACS impacts upon the information system's potential to contribute to workflow execution. This is a somewhat radical departure from Kaplan and Norton's (1992) original four perspectives. This perspective provides an aggregated view of these issues and asks, "Does the IS assist in the execution of the workflow?" This change was made to reflect the fact that PACS typically does not stand alone. Rather, in its relationship with existing systems, it can be integrated to streamline processes, or it can slow processes through parallel systems.

Figure 2. PACS-BSC model



These perspectives indicate another radical deviation from Kaplan and Norton's (1992) original BSC in which the perspectives indicated performance measures. While performance measures with their essentially qualitative implications are relevant to financial and corporate internal business process perspectives, it has been found that they may not be adequate for some enterprisewide applications (Goldszal, Bleshman & Bryan, 2004). When it comes to hospital goals of clinical excellence, the outcomes may be difficult to quantify because they have aspects that are essentially qualitative. It is hard to measure the intangible value created by PACS to patients, radiologists, and clinicians in terms of quantitative measures. It requires a broader understanding.

Therefore, the perspectives outlined here should be considered as incorporating qualitative themes rather than performance measures. This makes it necessary to have a qualitative method for understanding the values of the outcomes associated with many elements of the perspectives.

FUTURE TRENDS

This chapter highlights an important issue in health that will become increasingly important in the future: the multiple goals of health care providers requires a context-specific concept of strategy and appropriate evaluation techniques. The finding that fundamentally BSC is appropriate as a starting point, although not adequate to capture the complexity of public hospitals' multiple goals, suggests that conventional models from business system evaluation may be able to be modified meaningfully and valuably for health. This is good news, as it avoids the need to start from scratch in order to avoid the misleading application of "off the shelf" conventional models that are inadequate to evaluate health contexts. The trend toward evaluation using appropriate models and techniques heightens the prospect of greater accountability and system success in health. With the growing trend to see health information systems as integral to combating ill health and breaking in ill health/poverty nexus in the least developed nations, the need for appropriate evaluation is pressing.

CONCLUSION

Given the four perspectives that were found to capture the strategy in the particular case at hand, the proposed BSC to evaluate the impact of a PACS implementation on workflow in a hospital are as follows:

1. Clinical business process perspective
2. Patient perspective
3. Quality and transparency perspective
4. Information system perspective

These have been identified as consistent with the hospital strategy, as indicated by the users of PACS, other hospital officials, and the documents reviewed. These measures are consistent with the literature on workflow associated with the patient's clinical journey. Taken together, these perspectives give the following framework: the PACS-BSC model as seen in Figure 2. The PACS-BSC model enables the holistic investigation of PACS impacts in that it captures the essential impacts on workflow.

The model demonstrates interdependency of the four perspectives and the relationship with the vision and strategy of a hospital. This is in accordance with Kaplan and Norton's (1992) original intentions in that the four perspectives complement each other, and there is not one perspective that alone represents strategy; yet it is a significantly improved approach for specifically evaluating PACS. Moreover, the PACS-BSC invites qualitative assessment that reflects the qualitative nature of not only the perspectives, but also of the strategy and vision of hospitals.

The Balanced Scorecard allows adaptation to the specific needs of hospitals in order to evaluate organizational performance and IT/IS implementations. Because the implementation process is undertaken for reasons consistent with IS and organizational strategy that is largely nonmarket, it should be evaluated in nonmarket terms. Clinical goals and not-for-profit components should therefore be primary when evaluating IT/IS implementations in hospitals.

Our approach is radical in that it is built on a consideration of the fundamentals of hospital strategy, and valuable in that it results in the PACS-BSC model. Consistent with the qualitative nature of major aspects of hospital strategy, the new model incorporates qualitative themes rather than quantitative performance measures. While changing the perspectives to fit the hospital context and changing the quantitative focus of the performance measures to qualitative themes may be radical, there is no change to the essential nature of the BSC model. Rather, the new model retains Kaplan and Norton's (1992) intention to evaluate outcomes from the perspective of the organization's strategy and to be flexible to whatever those outcomes and strategies may be. Thus, the PACS-BSC model is a holistic method to evaluate PACS impact on workflow that is relevant to a hospital's not-for-profit and clinical strategies.

Retaining the original robustness of the BSC with regard to a radical review of organizational strategy and to the evaluation of qualitative outcomes, the PACS-BSC will be used to evaluate the impacts of PACS on workflow within hospitals. The authors argue that valid outcomes require a radical review of models in terms of purpose, concepts, and structure, and are currently undertaking empirical research to test the new model.

The application of the BSC framework and theories of strategy to PACS within the hospital environment has enabled both a richer understanding of the workflow outcomes and a refinement of the BSC. The PACS-BSC, as developed, has enabled a more diverse understanding of outcomes from the staff's various perspectives. The use of qualitative interview methods in conjunction with quantitative survey methods has enabled an unearthing of outcomes as observed by the staff. Therefore, it is concluded that the academic tool kit with various theories and methods enables a clearer view of reality, while reality enables the appropriate refinement of theory.

REFERENCES

Andrews, K.R. (1987). *The concept of corporate strategy*. Richard D. Irwin, Inc.

Andriole, K. (2002). Productivity and cost assessment of computed radiography, digital radiography, and screen-film for outpatient chest examinations. *Journal of Digital Imaging, 15*(3), 161–169.

Curtright, J.W., Stolp-Smith, S.C., & Edell, E.S. (2000). Strategic performance management: Development of a performance measurement system at the Mayo Clinic. *Journal of Healthcare Management, 45*(1), 58–68.

Firth, L., & Francis, P. (2004). Understanding the lack of adoption of e-commerce in the health sector: The clinician's strategic perspective. Proceedings of the IADIS Virtual Conference, Lisbon, 12–15.

Goldszal, A., Bleshman, M.H., & Bryan, R.N. (2004). Financing a large-scale picture archival and communication system. *Acad Radiol, 11*, 96–102.

Gordon, D., & Geiger, G. (2001). Strategic management of an electronic patient record project using the balanced score card. *Journal of Health care Information Management, 13*(3), 113–123.

Kaplan, R.S., & Norton, D.P. (1992). The balanced scorecard—Measures that drive performance. *Harvard Business Review*, January-February, 71–79.

Liedtka, J.M. (1992). Formulating hospital strategy: Moving beyond a market mentality. *Health Care Management Review*, Winter (17), 21–26.

McFarlan, F., McKenney, J.L., & Pyburn, P. (1983). The information archipelage: Plotting a course. *61*(1), 145–156.

Mooraj, S., Oyon, D., & Mastrettler, D. (1999). The BSC: A necessary good or an unnecessary evil? *European Management Journal, 17*(5), 481-491.

Peer, S., Peer, R., Walcher, M., Pohl, M., & Jaschke, W. (1999). Comparative reject analysis in conventional film screen and digital storage phosphor radiography. *EMJ Radiology*, *9*, 1693-1696.

Porter, M.E. (1996). What is strategy? *Harvard Business Review, 74*(6), 61–78.

Schriefer, J., Urden, D., & Rogers, S. (1997). Report cards: Tools for managing pathways and outcomes. *Outcomes Management for Nursing Practice, 1*, 14–19.

Strassman. (1990). *The business value of computers*. New Canaan, CT: Information Economics Press.

Willcocks, L., Petherbridge, P., & Olsen, N.A. (2001). *Making IT count: Strategy, delivery, infrastructure*. Oxford, UK: Butterworth-Heinemann.

KEY TERMS

Balanced Scorecard (BSC): BSC is a framework that measures organizational performance along various perspectives in accordance with the organizational strategy and vision.

Evaluation: The systematic acquisition and assessment of information using an overarching perspective to provide useful feedback about some sort of object/subject.

Hospital Strategy: An integrated set of actions with a focus on increasing the long-term prosperity and strength of hospitals relative to its environment and (strategic) goals.

Information Systems: A system that comprises people, machines, and/or methods organized to collect, process, transmit, and disseminate data that represent information.

Not for Profit: Not for profit is the concept of which the primary objective is to support some issue or matter of private interest or public concern for non-commercial purposes.

PACS: A Picture Archiving and Communication System (PACS) is an (workflow) integrated system that is designed to streamline the processes of acquiring, storing, distributing, and displaying medical images throughout a hospital's enterprise.

PACS-BSC Model: A Holistic Method. Based on the BSC, to evaluate PACS impact on workflow that is relevant to a hospital's not-for-profit and clinical strategies.

Strategy: A pattern of decision in a company that determines and reveals its objectives, purposes, or goals; produces the principal policies and plans for achieving those goals; and defines the range of business the company is to pursue, the kind of economic and human organization it is or intends to be, and the nature of the economic and noneconomic contribution it intends to make to its shareholders, employees, customers, and communities.

Parallel Architectures for MEDLINE Search

Rajendra V. Boppana
University of Texas at San Antonio, USA

Suresh Chalasani
University of Wisconsin at Parkside, USA

Bob Badgett
University of Texas Health Science Center at San Antonio, USA

Jacqueline A. Pugh
University of Texas Health Science Center at San Antonio, USA

INTRODUCTION

It is often extremely difficult in clinical practice to find answers in a timely manner to specific questions that arise in the management of patient health issues. The central repository of medical research is the MEDLINE database. The time needed to search and *review* results from MEDLINE may range from 20 minutes to 2½ hours (Lucas et al., 2004); this greatly exceeds the two minutes that clinicians typically have available to answer questions during clinical care (Ely et al., 1999). Consequently, clinicians rarely use MEDLINE (Ely et al., 1999); when they use MEDLINE, they do not use it well (Hersh & Hickam, 1998).

In 1997, the National Library of Medicine launched PubMed, which gave the first public and free access to MEDLINE with links to full texts of articles (Free Web-based access to NLM databases, 1997). To reduce the time needed to search and review results, SUMSearch was developed by Badgett, Paukert, and Levy (2001) to automatically examine the results from an initial MEDLINE search and revise queries multiple times to extract more relevant information. Thus, each query from a user may result in up to eight queries by SUMSearch of MEDLINE and other resources. An overview of SUMSearch and other search techniques may be found in the companion article (Badgett, Chalasani, Boppana & Pugh, 2007). However, owing to the restrictions on the access to MEDLINE by the National Library of Medicine, the refined queries need to be issued sequentially, and the rate of queries needs to be throttled. Due to Internet delays accumulating during multiple searches, SUMSearch's response time to a query can be as much as 30 seconds. This delay prevents adding to SUMSearch additional searches that are needed.

In this chapter, we describe a parallel architecture for MEDLINE database integrated with search refinement tools to facilitate accurate and fast responses to search requests by users. The proposed architecture, to be developed by the authors, will use low-cost, high-performance computing clusters consisting of Linux-based personal computers and workstations (i) to provide subsecond response times for individual searches and (ii) to support several concurrent queries from search refinement programs such as SUMSearch.

BACKGROUND

Several commercial companies have implemented fast search engines that run on a cluster of computers and sift through a large volume of content to find relevant information. For example, there are more than 20 billion Web pages registered with the Google search index as of 2006. Google uses sophisticated and proprietary techniques to find and return relevant Internet documents requested by the user. Fast searching can be accomplished by distributing the problem of searching among multiple computers that search simultaneously in various parts of data. Distributed processing and distributed databases have been significant research topics in past years. However, special techniques to optimize the response time of searching are required. For example, loading data to be searched into memory before any searching begins has been explored (Chalasani & Boppana, 2005). Constructing index trees based on keywords to speed up searches has also been explored by several researchers (Melnik, Raghavan, Yang & Garcia-Molina, 2001; Yu, Cuadrado, Ceglowski & Payne, 2003). Searching in the context of clinical setting was studied by Hersh and Hickam (1994), who showed that simple word searching is an effective means to search for medical literature. Hersh, et al. (2001) showed that Boolean searching and vector space searching are almost equally effective in the context of medical searches.

However, no comprehensive implementation that combines all these various techniques for fast searches has been reported in literature. Most such implementations are commercial and, hence, proprietary in nature. In this chapter, we will combine techniques such as (a) indexing large databases using keywords and content, (b) in-memory loading and searching of databases, (c) distributing content to be searched on multiple computers, and (d) increasing the precision of searches using finite refinements to achieve fast searching of MEDLINE data.

The rest of this chapter is organized as follows. Section 3 describes parallel architectures for the MEDLINE database. Section 4 discusses performance implications for the parallel MEDLINE implementation. Section 5 discusses directions for future research in this area. Section 6 concludes this chapter.

PARALLEL ARCHITECTURES FOR THE MEDLINE DATABASE

Proposed MEDLINE Database Architecture and Implementation

The proposed architecture for a parallel MEDLINE implementation is illustrated in Figure 1. This architecture exploits two types of parallelism:

- **Temporal parallelism.** The incoming requests are distributed equally among the application servers (Appservers). An incoming request is handled completely by a single Appserver.
- **Spatial parallelism.** The MEDLINE database is equally distributed among all database servers (DBservers). In response to an incoming search request, the documents corresponding to that search are retrieved from one or more DBservers.

In this architecture, the application server (Appserver) receives the search request. Each application server has the complete index tree for the MEDLINE database. The incoming requests are evenly distributed across application servers by the Gateway server (Sprayer machine). The application server quickly searches its index tree for the keywords contained in the request. The index tree search identifies the documents relevant for those keywords and the database servers (DBservers) on which those documents reside. The Appserver then requests the DBservers for the identified documents. DBservers return the documents to the Appserver, which then combines the results from various keywords and sends a response to the requestor. This procedure is indicated in Figure 2.

We need to address three critical issues for successful implementation of MEDLINE and to facilitate subsecond search times: (i) implementation of the search engine, (ii) implementation of the database, and (iii) automatic revision of searches. These are discussed further in the next three subsections.

Search Engine Implementation

To facilitate fast and multiple concurrent searches, index lists are often created for a database. An index list is similar to the index at the back of a textbook and indicates the IDs of records that contain a given word. For example, when a request for a list of documents containing the words "hypertension" and "diabetes" is submitted, the index lists for these two words will be examined to identify records that contain both words. Alternatively, a document vector that indicates words (in dictionary order) contained in a document can be used. Given the large vocabulary used in medical publications, the index list approach is preferable for this implementation. The records so identified will be retrieved from the DBservers and composed to generate a response. Figure 3 shows an example index list in the form of a tree search structure. Each keyword (or search term) is a node in this tree. Each node is associated with a list of database servers and the document IDs on that database server that point to relevant documents for that keyword. In the example shown in Figure 3, the relevant documents for the term "diabetes" are documents with IDs 2, 100, and 215 on DBserver 1, and documents with IDs 625, 901, and 1576 on DB server 5. Even though this example indicates a tree search mechanism (to simplify the discussion),

Figure 1. Parallel Architecture of the MEDLINE module; the actual hardware can be a simple fast Ethernet based Linux cluster

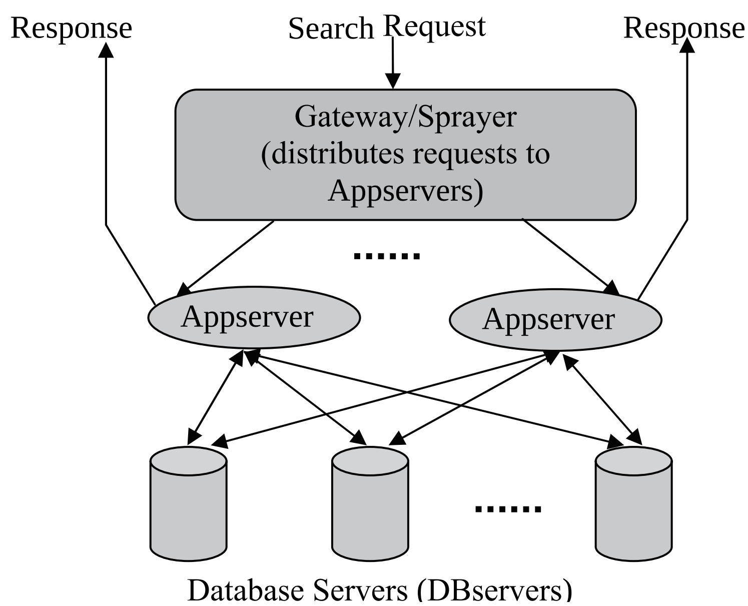


Figure 2. Process steps for handling a single MEDLINE search request

Process Steps for Servicing a MEDLINE Search Request by the Parallel Architecture
// Input to this process is the MEDLINE Search request including all the keywords.

Step 1: Gateway server (Sprayer machine) identifies an Appserver and forwards the search request to the Appserver.

Step 2: Appserver searches its copy of the index tree for each keyword and identifies the DBservers on which the corresponding documents reside. Appserver sends requests for those documents (it sends a unique request id, the keyword and the unique document ids for that keyword to the identified DBservers).

Step 3: Each DBserver which receives a request for the documents with document ids reads the corresponding documents from its memory and sends those requests to Appserver.

Step 4: The Appserver combines the resulting documents supplied by the DBservers and sorts the results by relevance. The Appserver then decides whether the search needs to be further refined; if so, it sends another set of retrieval requests to the DBservers. If the search does not need to be refined, it sends the results to the requestor.

the implementation in reality will follow a hash-map implementation so that searching for a keyword takes almost a constant amount of time, regardless of the search term.

Index lists will be created for the most frequently occurring words and phrases in the database. These index lists will be created at one time and stored (replicated) in the main memories of Appserver machines, which process the searches. We will write a multithreaded program to examine the index lists and identify the records that satisfy the search request. With multithreading, we can take advantage of multiple CPUs in each machine. To handle multiple concurrent searches, we can simply increase the number of Appservers.

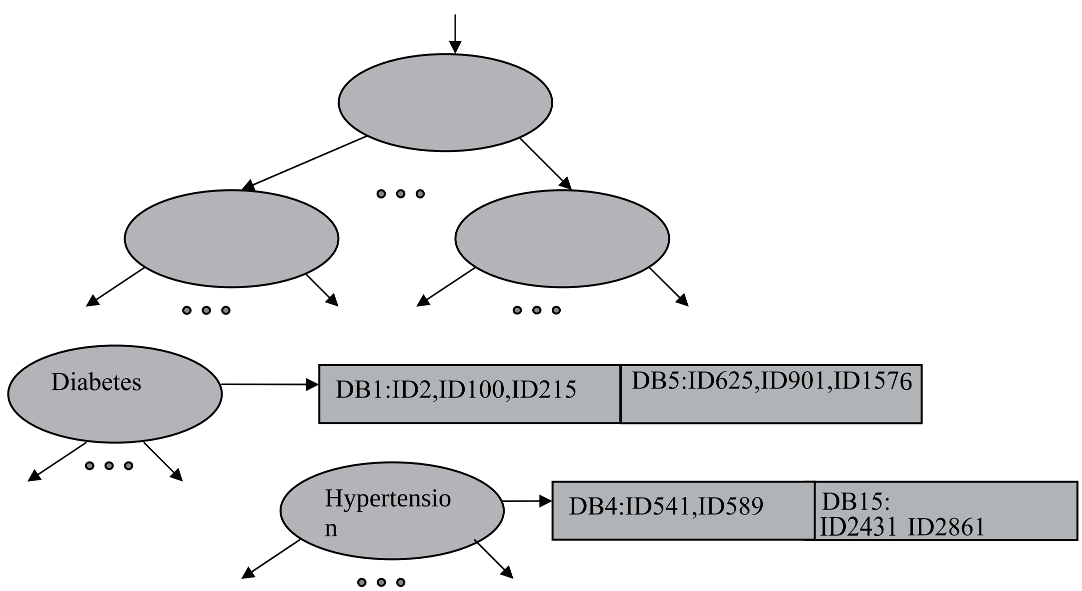
From the user's standpoint, a query is submitted to a Web server. For load balancing purposes, one machine can act as a sprayer or distributor of requests (see Figure 1) and distribute the request to an available Appserver. The Web server, which contains the controller module, acts as a conduit to pass the query request to a common gateway interface (cgi) program, programmed in C. This program will send the search request to an Appserver ensuring that all Appservers are evenly loaded with search requests. An Appserver in turn searches the index lists and retrieves appropriate records from DBservers. Then the Appserver examines the search results and combines it with results from other sources to format a response to the user. If necessary, to refine the search results, the Appserver will generate new search requests and process these new requests before it sends a final response to the user. This process is illustrated in Figure 2.

The size of the index list can fit easily in a few GBytes of memory. With 40 GB of data, it is reasonable to assume that there are a million documents in the MEDLINE database. Let us assume that there are 100,000 keywords used to search the database. Of these, 10,000 may occur frequently in the database. Such keywords are indexed by a binary string to indicate their presence in each document by a 1 or 0. Each such keyword requires a million bits, or 128 KB, of memory. In addition, the other less frequent keywords may occur in an average of 1,000 documents. In that case, each of the less frequent keywords requires 8,000 bytes (assuming each database record's uniquely located ID is 8 bytes long). The amount of memory required for all 100,000 keys is less than 800MB. For the 10,000 most frequent keywords, we will need 10,000 * 128KB = 1.28GB of memory. So the total memory requirements for keeping the index list in the main memory are slightly over 2GB. This is well within the scope of a moderately priced workstation.

Database Implementation

The database will be implemented using either the MySQL or PostGreSQL database engine. Both engines are high-performance engines that support multiprocessor hardware platforms. The MEDLINE database can be loaded into a sequential database from an XML format using software tools developed by Oliver, Bhalotia, Schwartz, Altman, and Hearst (2004). However, these tools need to be further modified to load the MEDLINE data into the parallel database architecture. The database can be searched using SQL syntax queries. To facilitate subsecond search times, we need to avoid searching the database from disk memory directly. To achieve this, we will split the database into nearly equal parts and allocate each part to a distinct machine (denoted DBserver). We will develop software that will let each DBserver (at boot time) read the database and store the records as Java objects in main memory. We will create a hash (mapping) mechanism to associate these Java objects with the unique IDs of the records they contain. To reduce the memory requirements, we will consider compressing infrequently used records. To improve

Figure 3. An example index list in the form of a search tree



performance, most of the data from the database will be loaded into the DBserver's memory. If there are 10 DBservers, each DBserver needs to load 4GB of data into its memory to load the complete 40GB MEDLINE database. DBservers with 8GB of memory can be purchased at a cost of $10,000 per server.

The MEDLINE database can be divided among the DBservers in two ways:

- **Strategy 1:** Distribute MEDLINE data equally among all DBservers.
- **Strategy 2:** Distribute MEDLINE data among DBservers so the overall load is approximately equally distributed at any given time.

Strategy 2 has the potential to improve the performance of our parallel implementation by distributing the load evenly among all DBservers. To implement this strategy, however, we need to estimate the probability with which each document in the MEDLINE database is accessed. For example, if document D_i is accessed with a probability of p_i, MEDLINE database documents need to be distributed among n DBservers such that the combined probability of all documents on any given DBserver is approximately 1/n. Probability p_i with which document D_i is accessed can be computed by looking at access trends of documents over an extended period of time. These trends likely can be predicted. For example, the publication date and nature of a document may predict frequency of access. Therefore, a database record that is a very old citation for a letter to the editor of a low-impact journal probably will be accessed infrequently. This strategy is difficult to implement in practice compared to Strategy 1 and will require quantifying correlated accesses of documents.

PERFORMANCE MODEL

As the number of users increases, more Web servers and Appservers can be added to handle multiple concurrent search requests. On the other hand, if the request load is not expected to be high, then we can combine the functionality of a Web server and Appserver into a single machine. We believe that our approach provides flexibility to scale up to handle larger databases and a larger request volume, if needed, and to scale down to facilitate a low-cost implementation if search time constraints are not stringent. We will use the Linux operating system with public domain MPI and OpenMP software packages to use multiple machines and multiple CPUs in each machine efficiently. For database implementation, we will use MySQL or PostGreSQL to implement DBservers. We will use Java and C to develop the search engine and other custom software as needed.

We will monitor the resulting MEDLINE Web service by observing its response times in response to queries from the search engine to the MEDLINE gateway. In addition to recording search times of the naturally occurring searches submitted by users, we also will submit automated requests every hour. The automated requests will be important during development when there are few requests for searches.

In addition, the static data from the MEDLINE database will be loaded into memory when the application starts and before any user request is processed. Our software will load the static data (data that does not change frequently) at the beginning of the application and store them in memory cache. As discussed in Section 2, we expect most of the data, including the MEDLINE data and the index trees, to reside completely in the memory; hence, no penalty for disk-reads is incurred in the parallel MEDLINE architecture.

Optimal Cluster Size

In this section, we present a simple analytical model to determine the optimal use of the cluster to speed up a single MEDLINE search. The response time of a query can be broken up into several components: (a) the time taken for a search request to reach an Appserver from a user machine; (b) the time taken for an AppServer to broadcast the search information to the Dbservers; (c) the time taken by DBservers to send data back to the AppServer; and (d) the time taken by the Appserver to format and send results back to the client machine. Of these, the time for tasks (a) and (b) are very small and are nearly constant for all searches; we can ignore these times in our model.

Let M be the average amount of data in bytes resulting from a single search. Let $M*t_a$, where t_a is the data format time per byte of data, be the amount of time it takes for an AppServer to format and send data to the client machine. Let $M*t_r$, where t_r is the retrieval time per byte of data, be the amount of time it takes to complete the search using a single processor. If n DBservers

P

are used to perform the search, then the search time is reduced to $(M*t_r)/n$. But this data need to be transferred from the DBservers to the AppServer. On average, each DBserver transmits M/n bytes of data.

The communication among servers in a cluster can be modeled as $t_s + m*t_b$, where t_s is the startup time required to format and prepare transmission of a message, t_b is the transmission time per byte of message, and m is the number of bytes transmitted (Grama, Gupta, Karypis & Kumar, 2003). So the total time taken to process a query, T_{search}, is the sum of the search time, the data transfer time, and the data format time.

$$T_{search} = (M*t_r)/n + n*[t_s + (M/n)*t_b] + M*t_a = (M*t_r)/n + n*t_s + M*t_b + M*t_a.$$

To minimize T_{search}, we need to differentiate the righthand side of this equation with respect to n, equate it to 0, and solve it for n.

$$-(M*t_r)/n^2 + t_s = 0$$

So, $n = (M*t_r/t_s)^{1/2}$ gives the optimal number of DBservers to be used to satisfy a single query.

The message startup time and computational time per unit data are dependent on the technology used. With fast Ethernet and standard Linux network protocol stack with parallel processing middleware such as MPI (message passing interface) (Gropp, Lusk & Skjellum, 1999), the startup time is about 100 microseconds. The retrieval time can be 100 nanoseconds (sufficient to execute about a thousand CPU instructions) or more per byte of data. If the average amount of data produced in a search is 100 Kbytes, then the optimal number of DBservers to be used is at least

$$(10^5 * 10^{-7}/10^{-4})^{1/2} = (100)^{1/2} = 10.$$

If the message startup overhead is reduced, for example, using more expensive Infiniband or Myrinet interconnects with optimized network software (Liu et al., 2003), then more DBservers can be used efficiently. On the other hand, if several queries need to be handled simultaneously, then more queries can be completed per unit of time if the number of DBservers is allocated proportionately to the amount of data to be generated by each query. We will incorporate this model to dynamically vary the number of DBservers used for efficient and fast completion of search requests.

FUTURE TRENDS

We intend to provide access to the MEDLINE database via the campus intranet initially and via the Web eventually to everyone else. The Internet presence includes a Web site that displays the user interface and also a Web service that exposes the MEDLINE database to remote searching by authorized collaborating Web sites. To facilitate easier user interface, we will incorporate user selectable clinical filters developed by other researchers in the medical field. The user may also access the MeSH browser to look up the canonical search term. Finally, the user will have the option to turn off automation if the user wants more control over the search strategy and making revisions.

The Web site will direct the user's query to the database and will receive the results from the database in XML. Scripting at the Web site will transform the XML into html, which is returned as a Web page to the user. The Web page of results that the user receives contains links to the full texts of the articles at the publishers' Web sites. To facilitate automated, high-volume queries by licensed third parties, we also will provide a direct query mechanism without going through the Web interface. In response to a query to the Web service, the search results will be sent as XML documents to allow the remote Web site to insert the results into its Web pages directly.

CONCLUSION

This chapter presented a parallel architecture for speeding up MEDLINE searches. The parallel MEDLINE architecture is based on constructing index trees for the MEDLINE database and storing this index tree in the memory of application servers. The index tree is replicated among all application servers, and the incoming requests for MEDLINE searches are evenly distributed among all application servers. However, the MEDLINE database is distributed among all database servers (or DBservers). In response to a search request, the application server (or Appserver) searches the index tree for keywords and identifies the DBservers on which the corresponding documents reside. The Appserver sends a request to read these documents to the DBservers and gathers the resulting documents from the DBservers and compiles a response. In some cases, the Appserver may further refine the searches based on the documents retrieved. This chapter presented an analytical model on the number of database servers needed and indicated techniques for performance improvement.

ACKNOWLEDGMENT

Dr. Boppana's research has been supported by National Science Foundation (NSF) grants EIA-0117255 and ANI-0228927, and the San Antonio Life Sciences Institute grant 10001642. Dr. Badgett and Dr. Pugh's research is supported in part by the VERDICT Research Enhancement Award Program, South Texas Veterans Health Care System Department of Medicine, and the University of Texas Health Sciences Center. The views expressed in this chapter are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs. This research is also sponsored in part by a summer research grant to Suresh Chalasani from the University of Wisconsin system.

REFERENCES

Badgett, R.G., Chalasani, S., Boppana, R.V., & Pugh, J.A. (2007). Medical search engines: Survey and issues. In N. Wickramsinghe, & E. Geisler (Eds.), *Encyclopedia of healthcare information systems*. Hershey, PA:IDEA Group Inc., 2007.

Badgett, R.G., Paukert, J.L., & Levy, L. (2001). The evolution of SUMsearch for teaching clinical informatics to third-year medical students. *Academic Medicine, 76*(5):541.

Chalasani, S., & Boppana, R. (2005). Software architectures for e-commerce computing systems with external hosting. *International Journal of Computers and Applications*, *27*(3), 190–198.

Ely, J.W., et al. (1999). Analysis of questions asked by family doctors regarding patient care. *British Medical Journal*, *319*(7206):358–361.

Free Web-based access to NLM databases. (1997). *NLM Technical Bulletin*, 296. National Institutes of Health: United States National Library of Medicine.

Grama, A., Gupta, A., Karypis, G., & Kumar, V. (2003). *Introduction to parallel computing* (2nd Ed.), Addison Wesley.

Gropp, W., Lusk, E., & Skjellum, A. (1999). Using MPI: portable parallel programming with the message passing interface (2nd Ed.). Boston, MA: MIT Press.

Hersh, W.R., & Hickam, D.H. (1998). How well do physicians use electronic information retrieval systems? A framework for investigation and systematic review. *Journal of the American Medical Association (JAMA)*, *280*(15):1347–1352.

Hersh, W., & Hickam, D. (1994). Use of a multi-application computer workstation in a clinical setting. *Bull Med Libr Assoc, 82*(4):382–389.

Hersh, W., et al. (2001). Challenging conventional assumptions of automated information retrieval with real users: Boolean searching and batch retrieval evaluations. *Information Processing & Management, 37*(3):383–402.

Liu, J., et al. (2003). Performance comparison of MPI implementations over InfiniBand, Myrinet and Quadrics. *ACM/IEEE Supercomputing*, 58.

Lucas, B.P., et al. (2004). The impact of evidence on physicians' inpatient treatment decisions. *Journal of General Internal Medicine*, *19*(5p1):402–409.

Melnik, S., Raghavan, S., Yang, B., & Garcia-Molina, H. (2001). Building a distributed full-text index for the Web. *ACM Transactions on Information Systems, 19*(3), 217–241. Retrieved from http://www-db.stanford.edu/~melnik/pub/melnik_TOIS01.pdf

Oliver, D.E., Bhalotia, G., Schwartz, A.S., Altman, R.B., & Hearst, M.A. (2004). Tools for loading MEDLINE into a relational database. *BMC Bioinformatics*, *5*(1), 146.

Yu, C., Cuadrado, J., Ceglowski, M., & Payne, J.S. (2003). Patterns in unstructured data. Retrieved January 15, 2005, from http://javelina.cet.middlebury.edu/lsa/out/cover_page.htm

KEY TERMS

Cluster of Computers: Several personal computers or workstations interconnected by an Ethernet or similar local area network and having a common communication software such as MPI (message passing interface) to solve various parts of a problem concurrently.

Index List: A list that indicates IDs of records that contain a given keyword.

MEDLINE: Database of citations and links to much of the world's biomedical research. MEDLINE is maintained by the National Library of Medicine of the National Institutes of Health. More information is available at http://www.nlm.nih.gov/databases/databases_medline.html.

Parallel Computing: The execution of a single task using multiple processors in order to speed up the task. In parallel computing, a single task may be divided into multiple subtasks that are assigned to the available processors. The results from these subtasks often are merged together to complete the task.

Performance Model: An analytical model used to predict various performance metrics such as speedup of search request response time, number of computers needed for optimal search times, and so forth.

Search Engine: A special software designed to retrieve from a database the records relevant to the search terms given by the user.

Servers: A computer designed to accept and execute certain types of tasks. Types of servers used in this chapter include Web servers, which interface with user via a Web browser; application servers, which interpret the incoming search requests and dynamically construct the search results; and database servers, which retrieve records based on the IDs indicated by the application servers.

Patients and Physicians Faced to New Healthcare Technologies

Marie-Christine Thaize Challier
University of Grenoble, France

INTRODUCTION

It is increasingly recognized that, by introducing healthcare information systems, it may be possible to improve healthcare delivery. Yet, insufficient attention has been paid, in the context of the introduction and implementation of new technologies, to examine the issues concerning the reduction of the astonishing variations in medical (and surgical) practices across seemingly similar areas. This article studies the causes of the differences in healthcare practices by examining the agency relationship where the patients delegate decision making to the physicians. To do this, it first presents the consequences of two main informational issues: the asymmetric information, and the imperfect information. Under such informational inadequacies, practitioners could not act in the best interests of their patients, and unnecessary and inappropriate cares could appear. Second, this article derives from such informational problems the necessity of information and coordination systems to ensure a better healthcare delivery.

BACKGROUND

Health researchers identify amazing variations in medical and surgical practices across otherwise-similar market areas (Phelps, 1997, *Darmouth Atlas of Healthcare,* cited by Dranove, 2000). Small Area Variations (SAVs), defined as wide variations in per-capita rates of healthcare procedures among seemingly similar areas, have many sources. The reasons for these variations, as well as their magnitude, remain controversial. Possible sources of the variations, besides demographics and insurance coverage, are patient preferences, the demand inducement, as well as the role of the differences in the practice style of physicians.

1. A first explanation of SAVs is attributed to the *patient preference*; that is, to their power to increase the healthcare demand (Folland, Goodman, & Stano, 2001, pp. 219, 222; Phelps & Mooney, 1992).
2. A second explanation of such differences rests on the existence of an *informational problem* due to the "supplier-induced demand" (SID) (Pauly, 1988; Phelps, 1986). This phenomenon is a problem of *asymmetric information* occurring when physicians use their influence to make treatment recommendations and other clinical decisions that rise the costs, and are not in the best interest of their patients. In that case, *patient ignorance* provides physician financial incentives to increase healthcare supply, and to generate unnecessary or inappropriate care. There is thus a specific informational asymmetry, called "moral hazard," revealing the postcontractual opportunism that occurs when a party takes an inappropriate action or provides distorted information (Arrow, 1963; Milgrom & Roberts, 1992). In this context, the role of physicians' payments (in particular under their three major forms) can be examined. *(i)* Under fee-for-service (FFS) basis, each physician has a schedule of fees and is reimbursed for each unit of service. Benefit schedules indicate the maximum levels reimbursed by insurers for a specific service. However, given this reimbursement, physicians could prescribe inadequate services in order to increase their own incomes. FFS compensation provides no financial incentives for health suppliers to mitigate moral hazard. *(ii)* Physicians can receive salaries/fixed fees (i.e., capitation for each patient assigned). This capitation (fee per capita and per year) limits repetitive visits, testing, and referrals; moral hazard can be reduced. Physicians have no incentives to provide unnecessary care given that they will not receive any additional payment for any additional care. However, if capitation induces physicians to rein in the utilization of their own services, this prepayment system does not prevent moral hazard, in particular the

overuse of specialists. Nevertheless, this problem is reduced if capitation fee increases when primary care physicians bear the costs of their referrals (see e.g., Dranove, 2000, pp. 75–76). *(iii)* Other physicians gain a fixed salary: a prepaid fee whatever the inputs provided. Through such a prefixed remuneration (i.e., a salary whatever the number of patients—moral hazard can be avoided).

3. Yet, most of controversy has focused on a third explanation: the differences in the *practice style of physicians* and the question of unnecessary or inappropriate care, due to *imperfect information* (Folland et al., 2001). Practice styles often differ across markets but also within communities. Where Green and Becker (1994) argued that SAVs are not necessarily due to the practice style hypothesis, most researchers find that many SAVs are closely related to the degree of physician uncertainty concerning the diagnosis and treatment (Wennberg, 1984). Corroborating this idea of physicians' uncertainty, healthcare researchers showed that the practice styles of physicians underline their imperfect information about the technology of producing health (Folland & Stano, 1989). In other terms, physicians could be misinformed (Folland et al., 2001). Consequently, physicians' decision making is not always optimal even if they act as "perfect agents" (i.e., without voluntary misbehaviors). Finally, such informational problems create substantial quality losses and wastes through overuse, underuse, and misuse.

Faced to such problems, the growing recognition of a need for better data and understanding led to various efforts to improve the healthcare system. Most of these actions developed evidence-based medicine, and recommended the collection and evaluation of administrative data in clinical and surgical services. In order to implement a variety of practices to identify information about cost and quality based on evaluations of outcomes research, a new tool emerged in the United States, the so-called "utilization review." Utilization review agencies usually approve the proposed treatment plans, but also recommend lower cost alternatives and sometimes recommend more costly interventions (Dranove, 2000, pp. 82). Detractors argue that such practices reduce costs at the expense of quality and they accuse the utilization review agencies of using "simplified treatment criteria that ignore patients' idiosyncratic needs" (Dranove, 2000, pp. 82). Due to contested practice guidelines and standards to assess and assure the healthcare quality, utilization review was eliminated in late 1999.

REDUCING SAVs

How can one provide a better way to avoid the part of SAVs due to informational inadequacies? The first way is accurately to identify the patients' and physicians' lack of information, and then to eliminate or reduce them by appropriate incentives and tools. Let us examine the challenges information systems face to reduce SAVs. One do not forget that the physician-patient transaction is an agency relationship where one party (here the patient) relies on another party (here the physician) to make decisions. Physician-patient transaction reveals a complicated relationship where the emergence of contingencies leads the parties to find adaptations. Consequently, this introduces the possibility of opportunistic behaviors leading to inefficiency.

Reducing SAVs in the Case of Asymmetric Information

Two types of *asymmetric information* can cause SAVs.

1. An important type of informational inefficiency leading to SAVs is due to the so-called "*adverse selection*" problem. Adverse selection refers to the *precontractual opportunism* due to the fact that one party has a private (i.e., unverifiable) information before getting engaged in a contractual relationship, this information affecting outcomes (Milgrom & Roberts, 1992). For example, patients may have private information about other symptoms; physicians may have private information about the available treatments, the efficiency of alternative medicines.

   To reduce the problem of asymmetric information relative to the physicians, it is necessary to develop systems of coordination displaying *common knowledge* about the availability of treatments and the capacity of therapies or alternative medicines, among others. In order to deal with the asymmetric information relative to the patients, the latter will report more easily the truth about their (other) current own healthcare characteristics if the information system allows the provider to check the patient history.

2. The second source of inefficiency is the *moral hazard* problem. For example, after the treatment delivery, patients may take actions that affect the outcomes—such as not follow the prescribed treatment—or they may also have incentives to lie about their healthcare conditions by understating them; physicians may prescribe unnecessary tests and procedures in order to protect themselves from malpractice suits.

   To deal with moral hazard problem, effective information systems can give the following solutions: *(i)* increase the resources of monitoring and verification in order to catch inappropriate actions before their occurrence; *(ii)* provide incentives for good behavior through rewarding good outcomes—for example, assess, develop, and implement new payment schemes (in order to remunerate good advice, Dranove (2000, pp. 174) imagines a "fee-per-advice" system that would resemble existing fee-for-service payments as well as a "retainer" system that would resemble capitation). *(iii)* Moreover, because *asymmetric information* leads to conflicts of interest between patients and physicians, it is necessary to align the individual interests with the group interest. Indeed, fear of opportunism may jeopardize the players' trust and lead to imperfect commitment. It is thus a priority to restore trust in physician competence and trust in the ability of healthcare systems to coordinate the various resources. All this constitutes a part of the "information integrity and quality" developed by HCIS (see, e.g., Eder, 2000; Fadlalla & Wickramasinghe, 2004).

Reducing SAVs in the Case of Imperfect Information

Other informational problems occur without opportunistic behaviors.

1. The motive behind the delegation of authority between a patient and a physician is that the former is generally less well informed about some features or problems than the latter. The *patient lack of information* refers to the true patient healthcare, the available treatments, the most appropriate treatment, the need to carry out other tests, and the expected outcomes of the prescribed treatment, among others.
2. However, it is insightful to note that the physician him/herself can also ignore the same features. Such a *physician lack of information* is due to uncertainty and ignorance in diagnosis and treatment. In addition, the physician can also ignore the other patient's symptoms, the patient's well being after the therapy, and the availability of various treatments, and the patient's preferences about the desired form of medicine. Physicians' uncertainty and ignorance over the best medical practices can also lead to unnecessary care.

   The resolution of such wastes is crucial. Solutions can be provided through information systems that can: *(i)* identify inappropriate or unnecessary care (i.e., define and measure what could be an appropriate and necessary care); *(ii)* assess and develop the taking into account of patient preferences, including the traditional and alternative forms of holistic medicine (Ayurveda and Chinese medicines, homeopathy, and herbal medicine, among others); and *(iii)* examine the patient needs (i.e., idiosyncratic needs).

Principles and Aims

More broadly, to solve such informational problems, the main question is: how to assess the effectiveness of healthcare information systems? The answer depends on two levels of information (Milgrom & Roberts, 1992, chap. 4). In the case of healthcare delivery, the answer is the following. First, if all information is reported honestly and if the information process is perfect and costless, the system of coordination can achieve an efficient healthcare decision. If it is not the case, an effective system must study how much extra information or coordination the system requires. Second, under imperfect information (i.e., if some healthcare information is lacking or inappropriate) it is relevant to determine what is the loss of effectiveness of the system. Taking into account such aspects, it is worthwhile to add that an effective system should apply the six key aims defined in 2001 by the Committee on the Quality of Healthcare (cited e.g., by Wickramasinghe, 2007, pp. 10). Indeed, healthcare should be: (1) safe–avoiding injuries to patients from the care; (2) effective—avoiding underuse and overuse; (3) patient-centered—avoiding care centered on the interest of other people; (4) timely; (5) efficient—avoiding wastes; (6) equitable—avoiding personal or social exclusion. Moreover, in the context of the emerging area of the "information integrity," four key principles are significant for the implementation of an effective system. Clearly, "the information must: (1) meet the consumers information needs; (2) be the product of a well-defined information production process; (3) be managed by taking a life-cycle approach; and (4) be managed and continually assessed vis-à-vis the integrity of the processes and the resultant information" (Huang, Lee, & Wang, 1999; Wickramasinghe, 2007, pp. 9–10).

FUTURE TRENDS

Medical practices can use information systems in order to reduce SAVs. Nevertheless, the future lies in finding new ways to improve the relationship between physicians and patients, and to provide direction for coordinating the delivery of healthcare resources. This article has examined some of the issues that must be taken into account when studying the relationships between informational issues, medical practices, and SAVs. However, several questions need further development. First, the relationship between physicians and patients need to be scrutinize in order to improve better information systems. Successful studies will show how the optimal healthcare decision can be detected in an environment characterized by a growing number of diagnostics, tests, and therapies through the generation and implementation of sophisticated technologies. Second, the strengths and weaknesses of information systems need careful investigations. When medical information is missing or imprecise (i.e., under imperfect information), it is crucial to find the real information in order to implement the best treatment. This will ensure superior decision making and support quality in healthcare delivery.

CONCLUSION

The generation and implementation of new information and communication technologies is currently affecting healthcare delivery and, in particular, the physician-patient relationship. Achieving high levels of information and coordination is a permanent challenge. Some researchers have focused on the reasons for SAVs, and the issues and controversies presented in this article have developed the role of asymmetric and imperfect information in the context of physicians' practices in connection with the healthcare preferences and needs of the patients. Medical and administrative data (on patient history, treatment protocols, research finding on treatments, among others) settle the foundation for a better healthcare system. Although information systems are already effective, ongoing efforts to generate much more quality improvement and reduce SAVs through appropriate information systems will create substantial welfare gains for society.

REFERENCES

Arrow, K.J. (1963). Uncertainty and the welfare economics of medical care. *American Economic Review*, *53*, 941–973.

Dranove, D. (2000). *The economic evolution of American healthcare. From Marcus Welby to managed care*. Princeton University Press.

Dranove, D., & White, W.D. (1987). Agency and the organization of healthcare delivery. *Inquiry*, *24*, 405–415.

Eder, L.B. (2000). *Managing healthcare information systems with Web-enabled technologies*. Hershey, PA: Idea Group Publishing.

Fadlalla, A., & Wickramasinghe, N. (2004). An integrative framework for HIPAA-compliant I*IQ healthcare information systems. *International Journal of Healthcare Quality Insurance*, *17*(1), 65–74.

Folland, S., Goodman, A.C., & Stano, M. (2001). *The economics of health and healthcare*. New Jersey: Prentice Hall.

Folland, S.T., & Stano, M. (1989). Sources of small area variations in the use of medical care, *Journal of Health Economics*, *8*, 85–107.

Green, L.A., & Becker, M.P. (1994). Physician decision making and variation in hospital admission rates for suspected acute cardiac ischemia. *Medical Care*, *32*, 1086–1097.

Huang, K.-T., Lee, Y.W., & Wang, R.Y. (1999). *Quality information and knowledge*. New Jersey: Prentice-Hall.

Milgrom, P., & Roberts, J. (1992). *Economics, organization and management*. Prentice-Hall International Editions.

Pauly, M.V. (1988). Is medical care different? Old questions, new answers. *Journal of Health Politics, Policy and Law, 13*, 227–237.

Phelps, C.E. (1986). Induced demand–can we ever know its extent? *Journal of Health Economics, 5,* 355–365.

Phelps, C.E. (1997). *Health economics*. Reading, MA: Addison-Wesley.

Phelps, C.E., & Mooney, C. (1992). Correction ad update on "priority setting in medical technology assessment". *Medical Care, 30*, 744–751.

Wennberg, J.E. (1984). Dealing with medical practice variations: A proposal for action. *Health Affairs, 3,* 6–32.

Wickramasinghe, N. (2007). Patient centric healthcare information systems in the U.S. *Mimeography*. Center for the Management Medical Technology, Chicago.

KEY TERMS

Adverse Selection (or Antiselection): In economic theory, adverse selection refers to the precontractual opportunism that results from an asymmetric information. It occurs when one party has private information—about his/her own characteristics or on the environment—that can affect the outcomes of the contract.

Agency Relationship: Situation in which one party relies on another party to make decisions.

Asymmetric Information: This term refers to situations in which contractual parties have differing amount of relevant information. Such situations can be adverse selection situation or moral hazard situation.

Imperfect Information: This term refers to situations in which a contractual party ignores relevant information about the other party or the environment, or when he/she does not know the action of the other party.

Moral Hazard: In economic theory, moral hazard refers to the postcontractual opportunism that results from an asymmetric information. It occurs when the actions contractually required are not freely observable; for example, a party takes an inappropriate action or decision that can affect the outcomes of the contract.

Private Information: Relevant information known only by some parties involved in a contractual relationship.

Small Area Variations: The wide variations in the per capita rates of healthcare procedures across seemingly similar areas.

P

Patients on Weaning Trials Classified with Neural Networks and Feature Selection

B.F. Giraldo
Universitat Politècnica de Catalunya, Spain

C. Arizmendi
Universitat Politècnica de Catalunya, Spain

E. Romero
Universitat Politècnica de Catalunya, Spain

R. Alquezar
Universitat Politècnica de Catalunya, Spain

P. Caminal
Universitat Politècnica de Catalunya, Spain

S. Benito
Hospital de la Santa Creu i Sant Pau, Spain

ABSTRACT

One of the challenges in intensive care is the process of weaning from mechanical ventilation. We studied the differences in respiratory pattern variability between patients capable of maintaining spontaneous breathing during weaning trials, and patients that fail to maintain spontaneous breathing. In this work, neural networks were applied to study these differences. 64 patients from mechanical ventilation are studied: Group S with 32 patients with Successful trials, and Group F with 32 patients that Failed to maintain spontaneous breathing and were reconnected. A performance of 64.56% of well classified patients was obtained using a neural network trained with the whole set of 35 features. After the application of a feature selection procedure (backward selection) 84.25% was obtained using only eight of the 35 features.

INTRODUCTION

Neural networks are very sophisticated modelling techniques capable of modelling extremely complex functions (Bishop, 1995). A simple network has a feed-forward structure: signals flow from inputs, forwards through any hidden units, eventually reaching the output units. Such a structure has stable behaviour. Neural networks are applicable in every situation in which a relationship between the predictor variables (independents, inputs) and predicted variables (dependents, outputs) exists, even when that relationship is very complex and not easy to articulate in the usual terms of "correlations" or "differences between groups."

The weaning trial protocols are very important in patients under mechanical ventilation. A failed weaning trial is discomforting for the patient and may induce significant cardiopulmonary distress. When mechanical ventilation is discontinued, up to 25% of patients have respiratory distress severe enough to necessitate reinstitution of ventilatory support (Tobin, 2001).

The respiratory pattern describes the mechanical function of the pulmonary system, and can be characterized by the following time series: inspiratory time (T_I), expiratory time (T_E), breath duration (T_{Tot}), tidal volume (V_T), fractional inspiratory time (T_I/T_{Tot}), mean inspiratory flow (V_T/T_I), and frequency-tidal volume ratio (f/V_T).

The variability of breathing pattern is nonrandom, and may be explained either by a central neural mechanism or by instability in the chemical feedback loops (Benchetrit, 2000). The study of variability in the respiratory pattern has been discussed in (Bruce, 1996; Caminal, Domingo, Giraldo, Vallverdú, Benito, Vázquez, & Kaplan, 2004; Khoo, 2000; Tobin, Mador, Guenter, Lodato, & Sackner, 1988).

The aim of the present work is the analysis with neural networks of the respiratory pattern variability in patients during weaning trials, in order to find differences between patients capable of maintaining spontaneous breathing and patients that fail to maintain spontaneous breathing.

As in many real situations, the suitable variables that describe the problem are partially unknown. When irrelevant variables are present, there may be many different models able to fit the data. But only some of them (those that do not use irrelevant variables) will lead to good generalization performance on unseen examples. However, in general it is not possible to control that irrelevant variables are not used during the training phase to learn the training set. This problem is shared by all modeling techniques, including neural networks. In this situation, feature selection techniques allow to search for a good subset of input features.

ANALYZED DATA

Respiratory flow was measured in 64 patients on weaning trials from mechanical ventilation (WEANDB data base). These patients were recorded in the Departments of Intensive Care Medicine at Santa Creu i Sant Pau Hospital and Getafe Hospital, according to a protocol approved by the local ethic committees. Using clinical criteria based on the T-tube test, the patients were classified into two groups: Group S (GS), 32 patients with Successful trials after 30 minutes; and Group F (GF), 32 patients that Failed to maintain spontaneous breathing and were reconnected after 30 minutes of weaning trials.

Respiratory flow was obtained using a pneumotachograph connected to an endotraqueal tube. The signal was recorded at a sampling frequency of 250 Hz during 30 minutes.

From each recorded signal the aforementioned time series were obtained: inspiratory time (T_I), expiratory time (T_E), breath duration (T_{Tot}), tidal volume (V_T), fractional inspiratory time (T_I/T_{Tot}), mean inspiratory flow (V_T/T_I), and frequency-tidal volume ratio (f/V_T).

METHODOLOGY

Data Preprocessing

Each one of the seven time series was processed by moving a Running Window (RW), with a width range from 3 to 100, consisting of several consecutive breath

Figure 1. Representation of 35 features for each patients obtained by Running Window (RW) with seven respiratory time series (TI, TE, TTot, VT, TI/TTot, VT/TI and f/VT and five statistics (mean, standard deviation, skewness, kurtosis and interquartile range) applied

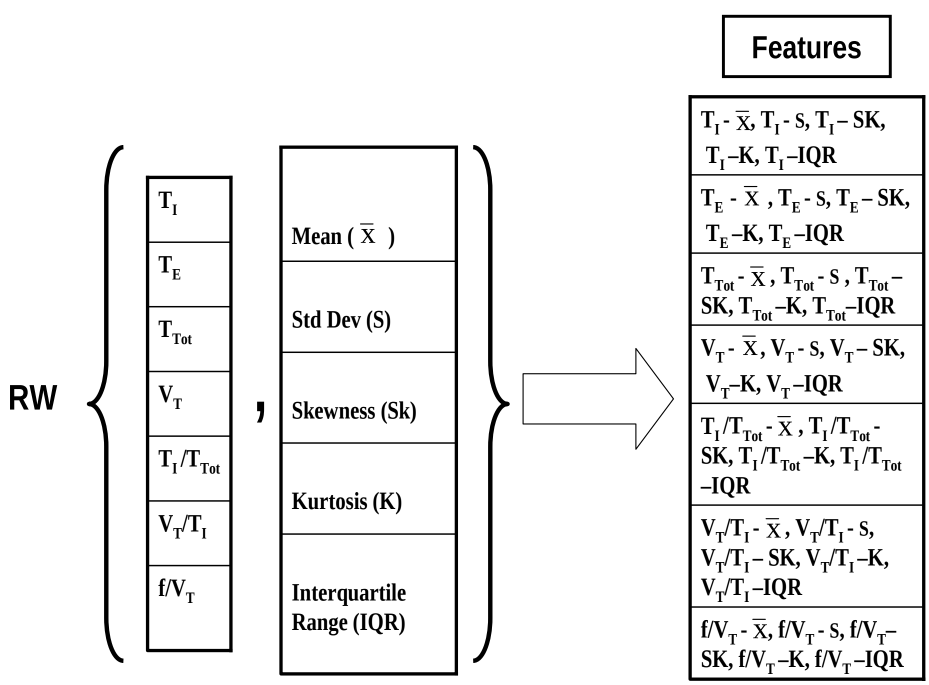
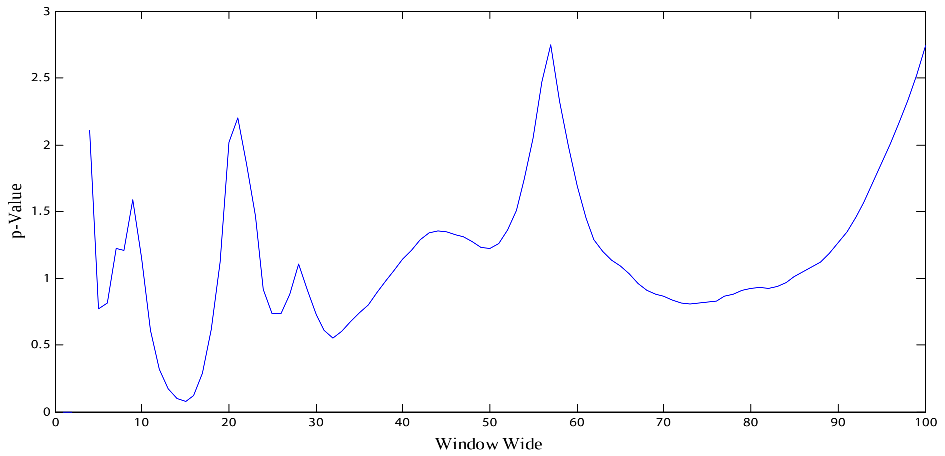


Figure 2. p-values total for all statistics parameters in all time series and for each window width. The best width was 15 (p < 0.001).



cycles. Five statistics were calculated for each window: the mean ($\overline{x}$), standard deviation (*S*), kurtosis (*K*), skewness (*Sk*), and interquartile range (*IQR*) of the value. In this way, 35 new time series were obtained for each patient (Figure 1), that are the features of the system. The optimal width of the RW was selected in the range from three to 100 by using a Mann-Whitney Test.

Figure 2 shows the total *p*-values for all statistics parameters in all time series and for each window width. The best width was 15, with $p < 0.001$ in all cases. Finally, this RW width was used to compute the 35 variables for each patient.

Next, with the RW width selected, the data of each patient were analyzed independently by applying a *k*-means clustering algorithm, which automatically determines the best number of clusters for all patients. The data of each patient are partitioned in *k* clusters mutually excluding. The partitions are defined by their centroides in each cluster, which represent the minimal distance between each data and this point. This cluster it must have a good cohesion (low intra-cluster variance). In this study, for all patients, there was a main cluster containing the most part of the patterns (i.e. data points). Table 1 shows the mean percentage of data points, represented by the main cluster for Group S and Group F.

Hence, this result was exploited to perform a dramatic data reduction, by which a single pattern of 35 features was associated with each patient. This pattern was computed as the mean value of the data points in the main (largest) conglomerate of the patient, using the *k*-means cluster algorithm.

As result of the data preprocessing step, the data set to be classified by the neural network included 64 patterns, one for each patient, 32 from Group S, and 32 from Group F.

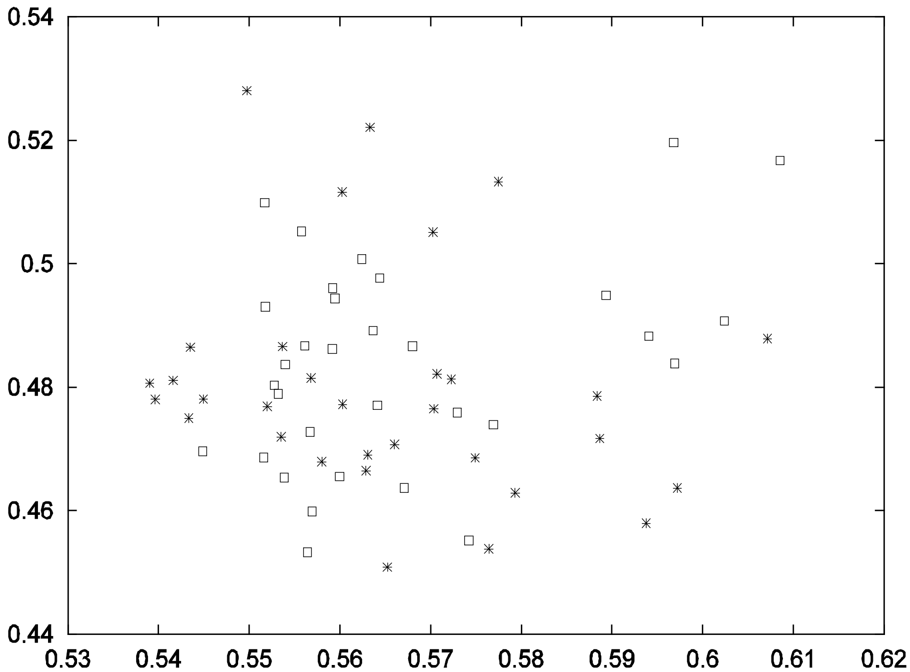
Data Visualization

In order to visualize the data, the 64 patterns were projected onto the plane with the Sammon mapping (Fiesler & Beale, 1997; Mao & Jain, 1995). Sammon mapping is a procedure designed to project high dimensional data onto low dimensional spaces preserving the original structure of the data. In particular, every point in the original space is associated to a point in the projected one, so that the distances between the original points are preserved in the projected space as much as possible. Figure 3 shows the results of the projection in this study (Sammon error = 0.053). As it can be seen,

Table 1. Percentage represented by the main cluster

	Mean ± SD (%)
Group S	
	96,081 ± 3,84
Group F	
	96,707 ± 3,58

Figure 3. Projection of the original data onto the plane with the Sammon mapping; different symbols indicate points of different classes



the data is not easily separable, and there are examples of different classes that are very near between them. This makes the data difficult to learn.

Neural Networks and Feature Selection

Neural networks are used for a wide variety of pattern classification tasks (Rumelhart, McClelland, & the PDP Research Group, 1986). A pattern is represented by a number of features that form a *d*-dimensional feature vector **x** within an input space $X \subseteq R^d$. A classifier therefore realizes a mapping from input space X to a finite set of classes $C = \{1,...,l\}$. A neural network is trained to perform a classification task from a set of training examples $S = \{(\mathbf{x}^{\mu}, t^{\mu}), \mu = 1,..., M\}$ using a supervised learning algorithm. The training set S consists of M feature vectors $\mathbf{x}^{\mu} \in R^d$ each labeled with a class membership $t^{\mu} \in C$. The network typically has as many outputs as classes and the target labels are translated into *l*-dimensional target vectors following a local unary representation. During the training phase the network parameters are adapted to approximate this mapping as accurately as possible (unless some technique, such as early stopping, is applied to avoid overfitting). In the test phase, an unlabeled feature vector $\mathbf{x} \in R^d$ is presented to the trained network, and the network outputs provide an estimation of the *a-posteriori* class probabilities for the input **x**, from which a classification decision is made, usually an assignment to the class with maximum *a-posteriori* probability (Bishop, 1995).

In this work, we used feed-forward two-layer perceptron architecture (i.e., one hidden layer of neurons and an output layer). For processing the full feature vectors, the networks consisted of 35 inputs, *n* hidden units and two output units, where *n* took different values from five to 30. Sine and logistic functions were used as activation functions in the hidden layer and the output layer, respectively. The networks were trained with backpropagation. A modified version of the Parallel Distributed Processing (PDP) simulator of Rumelhart and McClelland (Sopena, Romero, & Alquezar, 1999) was employed for the experiments. The weights were initially assigned randomly using a uniform distribution in the interval (-*w*range/2, *w*range/2). Some range values were tried and set independently for the hidden and output layers. Learning rates were adjusted for every architecture and range of weights. The momentum parameter was set to 0, with a maximum number of 1000 training epochs for each run.

For each network tested, 25 independent runs of an eight-fold cross-validation (CV) procedure were carried out, so that many different data partitions were generated, each including 56 patterns in the training set, and eight patterns in the test set. In some cases, eight patterns not used in the test set were reserved as validation set for early stopping the training phase (Romero et al., 2003). In this case, an internal seven-fold CV was performed with the training set of the eight-fold CV, so that six folds were used to train, one fold to validate and the remaining fold to test the model.

The network chosen at the end of the training phase was either the one that yielded the best classification result on the validation set among the networks obtained after each training epoch (when a validation set was used), or the one that yielded the best result on the training set (otherwise). Then, the chosen networks were evaluated on the corresponding test sets, and the test results were averaged for the different partitions and runs.

First, the experiments with the whole set of 35 attributes were carried out. As previously mentioned, 25 independent runs of an eight-fold CV (without validation set) were performed. After the experiments with the whole set of 35 features, we performed similar cross-validation experiments with different subsets of features, forming groups of five statistics related to each one of the seven time series collected. The number of inputs of the networks was set in this case to the dimension of the feature subset.

Finally, in order to determine the best possible subset of input features, feature selection was performed by carrying out several runs of a sequential backward selection method. Features were eliminated one by one, and the network was retrained each time a feature was temporarily removed. Again, eight-fold CV was applied and validation sets were used for early stopping and for selecting the features to remove. Once identified the feature subset providing the best result, 25 runs of an eight-fold CV (without validation set) were performed to obtain the final average test classification result using the selected variables.

RESULTS

Using all the 35 features, the average correct classification rate in the test sets was 64.56%. This result was improved in the case of using some fixed groups of the five statistics obtained from the original time series (Table 2), obtaining a best average result of 68.62%, using only skewness and standard deviation variables. However, a much more significant improvement in the average test result was achieved when applying the sequential backward feature selection process (in which features were not grouped but removed independently). Table 3 shows the three best average results after backward selection, being 84.25%, the best average of well classified patterns in the test set using only eight of the 35 features.

DISCUSSION AND CONCLUSION

Neural networks are very sophisticated modeling techniques capable of modeling extremely complex functions and feature selection techniques allow searching a good subset of input features.

Neural networks have been applied in order to analyze the respiratory pattern variability in patients during weaning trials, in order to find differences between patients capable of maintaining spontaneous

Table 2. Test classification results for several groups of selected variables using combinations of five statistics: Mean, standard deviation, skewness, kurtosis and interquartile range

Feature subsets	***Recall***		***Test result***
	GS	***GF***	***± SD***
Skewness (*Sk*) and Standard Deviation (*S*)	**69.75**%	**67.50**%	68.62% ±1.12%
Kurtosis (*K*), Skewness (*Sk*) and Interquartile Range (*IQR*)	65.75%	70.75%	68.25% ±1.21%
Standard Deviation (*S*), Skewness (*Sk*) and Interquartile Range (*IQR*)	67.75%	67.50%	67.62% ±1.24%
All 35 features	64.12%	65.00%	64.56% ±1.18%

Table 3. Test classification results for the best sets of selected variables obtained after sequential backward feature selection

Feature subsets obtained after sequential backward selection	***Recall***		***Test result ± SD***
	GS	***GF***	
$Sk(V_T/T_I), K(V_T), Sk(T_I/T_{Tot}), K(T_I), Sk(T_I), IQR(f/V_T), \overline{X}(V_T), IQR(T_{Tot})$	76.75%	91.75%	84.25% ±0.81%
$S(V_T/T_I), Sk(f/V_T), S(T_E), \overline{X}(T_I), IQR(f/V_T), Sk(V_T/T_I), IQR(T_I), Sk(T_I), K(T_{Tot}), K(V_T), Sk(T_I/T_{Tot}), IQR(T_{Tot}), \overline{X}(V_T/T_I), \overline{X}(T_I/T_{Tot})$	79.12%	83.00%	81.06% ±1.13%
$\overline{X}(f/V_T), IQR(f/V_T), Sk(V_T/T_I), Sk(V_T), IQR(V_T/T_I), IQR(T_I/T_{Tot}), \overline{X}(T_I/T_{Tot})$	79.75%	80.62%	80.19% ±0.87%

breathing and patients that fail to maintain spontaneous breathing.

The networks trained with the whole set of 35 features achieved a performance of 64.56%. After the application of feature selection, 84.25% was obtained using only eight of the 35 features. Although the groups were balanced, a greater facility of learning and generalization was observed in the Group F with 91.75% of successes compared with the Group S with 76.75%.

It is necessary to make a model with the most important characteristics and to prove the generalization of our system. Additionally, it could be interesting to increase the number of patterns used for training, in order to improve the generalization error, assuring the presence of the subgroup with the most important characteristics.

The optimal wide of the RW was computed with the complete set of characteristics, when selected characteristics are modified, this optimal wide could change.

Given the clustering results, it would be important to carry out a study to obtain the indexes that allowed us to decrease the temporary registry of data.

ACKNOWLEDGMENT

The authors would like to thank to Dr. A. Ballestero of the *Hospital Universitario de Getafe*, Getafe, Spain, by his collaboration in the signal database acquisition.

This work was supported in part by Ministerio de Educación y Ciencia, under grants TEC2007-63637, TEC2007-68076-C02-01 from the Spanish Government.

This work was produced in affiliation with the CIBER de Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN) and the Institut de Bioenginyeria de Catalunya (IBEC), Spain.

REFERENCES

Benchetrit, G. (2000). Breathing pattern in humans: Diversity and individuality. *Respiration Physiology*, *122*, 123–129.

Bishop, M. (1995). *Neural networks for pattern recognition*. Oxford University Press.

Bruce, E.N. (1996). Measures of respiratory pattern variability. In *Bioengineering Approaches to Pulmonary Physiology and Medicine* (pp. 149–160). Plenum Press.

Caminal, P., Domingo, L., Giraldo, B.F., Vallverdú, M., Benito, S., Vázquez, G., & Kaplan, D. (2004). Variability analysis of the respiratory volume based on nonlinear prediction methods. *Medical & Biological Engineering & Computing*, *42*, 86–91.

Fiesler, E., & Beale, R. (Eds.) (1997). *Handbook of neural computation*. IOP Publishing Ltd. and Oxford University Press.

Khoo, M.C. (2000). Determinants of ventilatory instability and variability. *Respiratory Physiology*, *122*, 167–182.

Mao, J., & Jain, A.K. (1995). Artificial neural networks for feature extraction and multivariate data projection. *IEEE Transactions on Neural Networks*, *6*(2), 296–317.

Romero, E., Sopena, J.M., Navarrete, G., & Alquezar, R. (2003). Feature selection forcing overtraining may help to improve performance. In *Proceedings of the International Joint Conference on Neural Networks, (IJCNN-2003)* (Vol.3, pp. 2181–2186). Portland, OR.

Rumelhart, D.E., McClelland, J.L., & the PDP Research Group (Eds.) (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. MIT Press.

Sopena, J.M., Romero, E., & Alquezar, R. (1999). Neural networks with periodic and monotonic activation functions: A comparative study in classification problems. In *Proceedings of the International Conference on Artificial Neural Networks, (ICANN99)* (pp. 323–328).

Tobin, M.J. (2001). Advances in mechanical ventilation. *New England Journal of Medicine, 344*(26), 1986–1996.

Tobin, M.J., Mador, M.J., Guenter, S.M., Lodato, R.F., & Sackner, M.A. (1988). Variability of resting respiratory center drive and timing in healthy subjects. *Journal of Applied Physiology, 65*, 309–317.

KEY TERMS

Data Visualization: Technique of looking at data to identify like and unlike groups of data points.

Neural Networks: Form of artificial intelligence.

Respiratory Pattern: Breathing pattern.

Spontaneous Breathing: Normal/natural breathing.

Planning and Control and the Use of Information Technology in Mental Healthcare Organizations

I.J. Baars
Maastricht University, The Netherlands

G.G. Van Merode
Board of Maastricht University Medical Center & Maastricht University, The Netherlands

INTRODUCTION

Demand for mental healthcare increases. Simultaneously, the need for more patient oriented processes increases and the market develops towards more competition among providers and organizations. As a result of these developments, mental healthcare organizations are becoming more aware of efficiency and effectiveness. Often, they choose to transform to more process oriented organizations, which require changes in planning and control systems and information technology (IT). However, little is known about the required planning and control systems and IT for mental healthcare.

We argue that IT for planning and control of mental healthcare organization needs to be adaptive and support short term planning. IT has to be adaptive to be able to support first and second order control which is needed for planning and control of mental healthcare processes. Short term planning or reactivity is needed to deal with stochasticity and variability as present in mental healthcare. These subjects are further described in the background.

This article reports the results of two studies on the use of standard care processes and IT for planning and control of mental healthcare processes. The results give insight in the needed functionalities of IT and planning and control of mental healthcare processes. The first study is a case study in a center for multidisciplinary (mental) youth care. This center implemented care programs and an automated planning tool. We studied the success of this implementation and particularly the fit between the care programs and the planning tool.

In the second study we studied the characteristics of ambulant mental healthcare processes and the actual and preferable use of planning and control models and IT.

BACKGROUND

Mental healthcare is often multidisciplinary and includes several professionals, disciplines, and departments within one or more organizations which all need to be planned and controlled. Consequently, the object of control is mainly professionals and patients, but also resources like rooms. We define planning as the determination of what should be done and control as the process that assures that the planned results are obtained (Van Merode, Groothuis, & Hasman, 2004).

According to Hofstede (1981), the way nonprofit organizations, such as mental healthcare organizations, can be planned and controlled depends on the type of processes. The type of process can be determined by answering the following questions: is the output measurable? Are the objectives unambiguous? Are the effects of management interventions known? And, can the activities be repeated? The type of processes determines the control model and instruments (e.g., protocols, case management and budgeting) that can be applied.

Hofstede (1981) defines six different control models, as shown in Table 1. The more standardized, well-defined, and structured the processes are, the more routine control can be used.

A mental healthcare organization consists of various processes. These processes are possibly different and thus need different control models. To select a control model that best suits a situation, processes have to be analyzed. Especially the distinction between routine and less routine processes is important here.

Routine processes can use standards and can be controlled by routine control or, when the activities cannot be repeated, by expert control. For routine control, and marginally for expert control and trial-and-error control, models that compare automatically what

Table 1. Control models

Control model condition	Unambiguous objectives	Measurable output	Known effects	Repetitive activities
Routine control	+	+	+	+
Expert control	+	+	+	-
Trial and error control	+	+	-	+
Intuitive control	+	+	-	-
Judgmental control	+	-	+/-	+/-
Political control	-	+/-	+/-	+/-

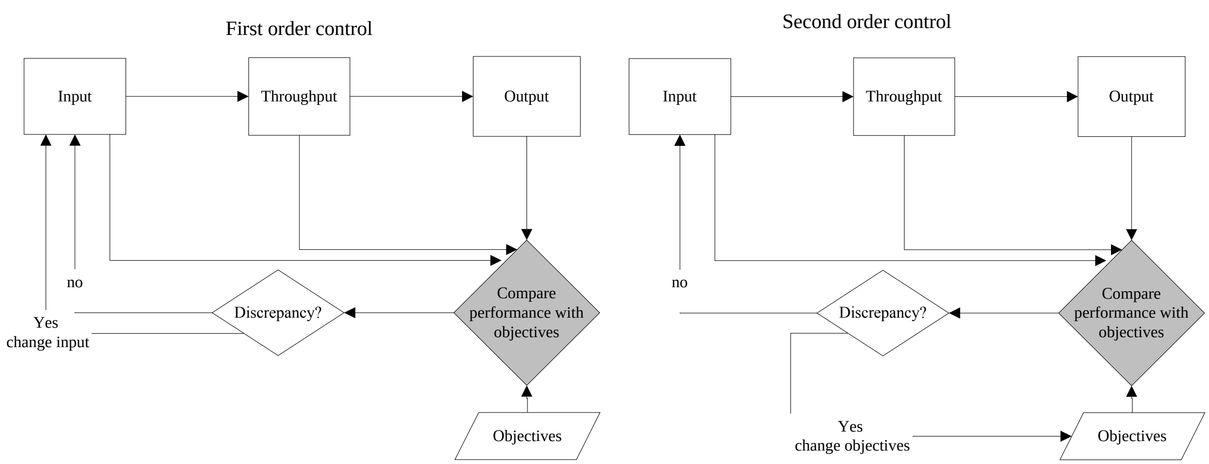
+ = condition is present
- = condition is not present

actually happens with the standards about what should happen apply. These models use feedback for control. The feedback from relevant indicators provides information that can be compared with targets. A number of care processes can be coordinated to a large extend by planning and control integrated in one system and by providing feedback from control to planning (Hofstede, 1981; Van Merode, Groothuis et al., 2004; Van Roth & Van Dierdonck, 1995). Moreover, for intuitive and judgmental and political control only vague models exist (Hofstede, 1981).

An example of control by feedback in mental healthcare is a protocol for the process of the treatment of depression. This protocol describes that the target of the indicator "number of sessions" is 15. If a patient receives 17 instead of the targeted 15 sessions, this is evaluated. Several actions can be undertaken to meet the target of 15. One such an action is adapting the input of the process so that the 15 sessions are not exceeded. This is first order control (Figure 1).

A problem with first order control can be that the aim becomes to keep the activity on target at any cost. To overcome this problem, a second order feedback loop that can periodically adjust the targets of the first order feedback loop is necessary (Figure 1). A second order feedback loop provides information (e.g., by new insights or outcome of evaluations) that enables decisions on the appropriateness of the target. By means of a second order feedback loop, the interventions of the first order loop can be overruled. With that, the targets of the process can be adapted so that the outcome better fits the targets. In our example of 15 sessions where the output did not meet the target of 15 sessions, the target of the standard can be changed to, for example, 17 sessions. An advantage of second order feedback is that it can cover more complex organizational control

Figure 1. First and second order control



situations. Furthermore, variability and stochasticity can be accounted for with a second order feedback loop.

Information Technology for Planning and Control

Planning and control by feedback on performances can be supported by IT. IT provides the backbone for collecting, compiling, and utilizing information on patients, activities, methods, costs, and results (Porter & Olmsted Teisberg, 2006). Because of the multidisciplinary care and the involvement of several professionals, disciplines, and departments that need to be planned and controlled, a central automated system for planning and control might be needed. In industry Enterprise Resources Planning systems are often used for this purpose. Enterprise Resource Planning (ERP) systems attempt to integrate all corporate information in one central database, and information can be retrieved from many different organizational positions and in principle they allow any organizational object to be made visible (Dechow & Mouritsen, 2005).

Likewise, for healthcare, ERP systems can integrate many functions such as patient scheduling, human resource management, workload forecasting, and management of workflow (Jenkins & Christenson, 2001). However, when implementing ERP systems in an organization, several implementation and structural problems can occur, for example, ERP may not fit the structure of the organization.

Implementation and structural problems may occur in healthcare because ERP systems require fixed, deterministic processes and ignore alternative processes (Van Merode, Groothuis et al., 2004). But in healthcare, stochasticity (e.g., patients not showing up) and variability (e.g., demands for care differ between several individual patients) exist. Because of the existence of stochasticity and variability, a more reactive decision making technology or short term planning systems as, for example, Advanced Resource Planning (ARP), better suit the less deterministic processes (Vandaele & De Boeck, 2003; Van Merode, Groothuis et al., 2004). Another possibility for planning and control in mental healthcare is to physically separate deterministic processes suitable for ERP from nondeterministic processes requiring a more reactive decision making, or short term planning, technology.

PLANNING AND CONTROL AND THE USE OF IT IN MENTAL HEALTHCARE ORGANIZATIONS

To study the possibilities of planning and control and the use in mental healthcare organizations, we performed two case studies. Here we present the results of these two case studies.

Case 1: Planning Tool and Care Programs

In the multidisciplinary youth (mental) healthcare center, processes were standardized by care programs. The aim of the care programs was to plan processes more efficient and effective, and also to improve the patient-oriented way of working. A care program is defined as a framework with which organizations, professionals, and patients should comply and it is used to get a patient-oriented organization of a well described target group. It integrates the activities between different disciplines, professionals, and departments and indicates the way the care should be given, by which professional, in which setting, and with what frequency (Berg, Schellekens et al., 2005).

Care programs consist of several standardized modules. The planning of a care program, that is, which modules will be performed, is done based on the demand of the patient. Further, for planning and control of the activities of a module, each module contains a Bill of Resources (BOR), which is described in the protocol of the module.

A BOR is derived from a Bill of Materials that is used by ERP systems. However, mental healthcare includes more than materials such as professionals and patients. Therefore, the use of a BOR is more appropriate for healthcare organizations (Van Roth & Van Dierdonck, 1995). In the center, a BOR defined the location, type of activity, professional, and patient.

Because of the multidisciplinary care, the coordination of the care programs was complex. Automated support might reduce this complexity. Therefore, IT, that is, the planning tool, was used for planning and control of activities of patients and professionals. The planning tool was a software application, intended to support working with care programs. The features of the care programs, such as the involvement of several professionals and departments and the standardization of modules, require certain functionalities of the planning tool. The planning tool contained the following functionalities: a BOR to plan activities, an event handler to react to ad hoc changes by revising timetables and re-arranging activities, an electronic organizer, and a control function for process control by first order feedback and second order feedback.

To test the functionalities in relation to the care programs, the planning tool was tested in a pilot. The evaluation of the pilot was done with a document analysis, interviews with all those involved, and analysis of e-mails to and from the planning tool helpdesk.

The results of the evaluation show that the fit between the standardized care programs and the planning tool was insufficient. The main cause was the noncompliance of the professionals with the standard modules because they feared inflexibility in performing their activities due to the standardization and first order control. However, the professionals did have some flexibility in planning. They could use the event handler to react to ad hoc situations and they had the possibility to deviate from the standards to react to variability which is second order control. But the results show that the flexibility possibilities were not known and not used by the professionals.

As the results illustrate, the characteristics of care processes and the functionalities of IT must fit to function. In this case, the IT demands a certain amount of standardization while in daily practice the professionals require flexibility in performing their activities. Therefore standardization and flexibility have to be balanced. Because the professionals did not comply with the standard modules of the care programs, and did not know of, or use, the possible flexibility, we do not know for sure that the care programs and the IT did not fit. It seems also due to implementation problems of the IT and care programs that professionals did not comply and did not have knowledge about the possibilities of second order control.

Professionals do need some planning freedom to be able to react to the present variability and stochasticity. A planning and control system, including the IT, need to support this in an efficient way. In the second case study, we studied the actual use and the possibilities of a planning and control model which is based on performance measurement and second order feedback.

Case 2: Optimizing Planning and Control in Ambulant Mental Healthcare Centers

Due to increasing competition in mental healthcare, planning and control based on performances is needed. Therefore routine control is necessary and processes have to fulfill the characteristics as described by Hofstede (1981). But as shown in the evaluation of the planning tool, also flexibility in planning is needed.

The case study of four ambulant mental healthcare centers presents information about characteristics of their processes and the planning and control models as used. These results are used for recommendations that can guide mental healthcare organizations in their planning and control and the use of IT. These are presented in the conclusion.

In this case study, we first described the present processes by using process mapping techniques. Next, we assessed these processes on the uncertainty of demand, supply, and the service process itself, complexity of coordination, and staff inflexibility. These three factors determine the efficiency and adaptation possibilities of a process (Van Merode, Molema, & Goldschmidt, 2004). These factors incorporate the conditions of Hofstede's (1981) model.

The results of this study showed that most of the processes were not planned and controlled in a consistence way. Three of the four ambulant mental healthcare centers hardly used performance indicators to monitor the processes. The uncertainty and complexity were high and instruments, like protocols or IT, to decrease the unnecessary uncertainty and complexity were hardly used. One center did use care programs. However, the opportunities for planning and control of these care programs were not used to their full extent. Besides, a planning and control system based on feedback did not exist. This center had the most extended IT software. Nevertheless, this software was not used to monitor the results of the processes with performance indicators. Therefore, planning and control based on feedback was not performed. IT use for planning and control was very minimal in all four centers.

FUTURE TRENDS

In mental healthcare organizations, we mainly see two developments. The first is that many mental healthcare organizations continue in their old ways in organizing their care, that is, that no standards or IT are used. However, a second development is that, because of growing attention for more efficiency in mental healthcare, more and more mental healthcare organizations standardize their processes by, for example, the introduction of care programs.

Standardization can be very fruitful in decreasing unnecessary uncertainty in care processes. However, mental healthcare organizations often do not change their organization structure simultaneously with the introduction of care programs. Additionally, IT with planning and control functionalities is underused. The use of IT for integrated planning and control will probably increase. But, what we observe in somatic healthcare is that (standard) ERP packages are implemented while they do not suit the situation. These ERP systems are often implemented without a change in organization structure. By the lack of the fit between the organization, processes, and ERP systems, this can result in many problems. New, adaptive, software is there but is not successful.

As we claimed in the introduction, first and second order control should be supported by IT. Therefore IT has to be adaptive and support short term planning or reactive decision making. However, mental healthcare organizations are not yet interested in this kind of technology. In the future, mental healthcare organizations consider IT more to profit from care programming and, as a result, work more efficient and effective.

CONCLUSION

Mental healthcare organizations are on their way to more efficiency in their processes; yet IT is still underused for planning and control.

As shown in the first case study, processes and IT must fit. In this center, the organization thought that planning and control was possible with routine control supported with information technology. The care programs met the conditions for first order control and second order control. However, in the pilot it did not work out as intended.

In the second study, we observe that almost no use is made of possibilities of standardization and feedback for control purposes. Moreover, IT is hardly used for planning and control purposes. As a result, the ambulant mental healthcare centers miss the possibility to direct the outcome of processes to a certain target, and efficiency is not optimal. As a result, the centers miss the opportunities to be competing on the (regulated) market.

The results emphasize the need to carefully consider process characteristics before introducing a planning and control system. Careful consideration is needed to avoid type I and type II errors. Type I errors occur when opportunities for routine planning and control (with standardization and IT) are not used by management. However, some processes are not deterministic and cannot be controlled with routine models. When they are controlled with routine models anyway, type II errors occur. To avoid either type I or type II errors, a fit between standardization and flexibility has to been found. Hofstede (1981) states that type I errors often occur in not-for-profit organizations because the concern for cost and effectiveness is often missing. However, mental healthcare is subject of social and political developments resulting in the need to be more aware of efficiency and effectiveness. Therefore we assume that type II errors may occur in future more often.

The results also show the necessity of balancing between process standardization and the flexibility of the working practice. Routine models can be applied for processes in mental healthcare in certain circumstances. Nevertheless, routine models need to use second order control because not all activities can be planned in advance due to variability and stochasticity. This balance between standardization and flexibility has to be considered carefully before fully implementing care programs and a central planning and control system integrated in IT. To be able to balance between flexibility and efficiency by means of care programs, the organization structure also has to be changed and IT has to be adaptive.

We recommend the following to develop more efficient processes in mental healthcare organizations. First, performance indicators should be developed to be able to compete in the (regulated) market. Simultaneously, time horizons should be distinguished in which planning and control on several different levels (e.g., establishing different care programs by board 3 years, planning of professional 6 weeks) can be defined. Next, processes need to be distinguished on uncertainty and complexity of the process. The difference in characteristics of a process (more or less routine) decides on the possibility of standardizing processes. After that, the processes that satisfy the possibilities for routine control need to be standardized and a BOR should be described. For the actual control on performances, a system based on first order control and second order control must be developed and adaptive IT that supports short term planning or is reactive is needed to support the planning and control.

The use of these recommendations helps in finding a balance between standardization, use of IT, and flexibility. In addition, IT interacts with the social system and the working practice and, therefore, it is necessary to tailor both. However, it is a process of trial and error and therefore requires adaptive IT.

REFERENCES

Berg, M., Schellekens, W., et al. (2005). Bridging the quality chasm: Integrating professional and organizational approaches to quality. *International Journal Quality of Healthcare, 17*(1), 75-82.

Dechow, N., & Mouritsen, J. (2005). Enterprise resource planning systems, management control and the quest for integration. *Accounting, Organizations and Society, 30,* 691-733.

Hofstede, G. (1981). Management control of not-for profit activities. *Accounting, Organizations and Society: An International Journal Devoted to the Behavioural Organizational & Social Aspects of Accounting, 6,* 193-211.

Jenkins, E.K., & Christenson, E. (2001). ERP (enterprise resource planning) systems can streamline healthcare business functions. Healthcare Financial Management, 55(5), 48-52.

Porter, M.E., & Olmsted Teisberg, E. (2006). Redefining healthcare. *Creating value-based competition on results.* Boston, MA: Harvard Business School Press.

Vandaele, N., & De Boeck, L. (2003). Advanced resource planning. *Robotics and Computer Integrated Manufacturing, 19,* 211-218.

Van Merode, G.G., Groothuis, S., & Hasman, A. (2004). Enterprise resource planning for hospitals. *International Journal of Medical Informatics, 73*(6), 493-501.

Van Merode, F., Molema, H., & Goldschmidt, H. (2004). GUM and six sigma approaches positioned as deterministic tools in quality target engineering. *Accreditation and Quality Assurance, 10,* 32-36.

Van Roth, A., & Van Dierdonck, R. (1995). Hospital resource planning: Concepts, feasibility and framework. *Production and operations management, 4*(1), 2-29.

KEY TERMS

Adaptive Information Technology: Information Technology that can be adapted to changing circumstances in finding a balance between processes and IT.

Care Program: A care program is a framework where organizations, professionals, and patients should comply with and is used for patient-centered organization of a well described target group. It integrates the activities between different disciplines, professionals, and departments (Berg, Schellekens et al., 2005).

Control: The process that assures that the planned results are obtained.

Enterprise Resource Planning System: Enterprise Resource Planning systems attempt to integrate all corporate information in one central database, and information can be retrieved from many different organizational positions and in principle they allow any organizational object to be made visible (Dechow & Mouritsen, 2005).

Feedback: Information about the output is fed back to the input.

Planning: Determination of what should be done.

Reactive Decision Making: Decision making based on reacting to unexpected situations due to variability and stochasticity.

Short Term Planning: Planning on short terms with data about monthly, daily, and hourly demand and meaning full statistical distributions to be able to deal with stochasticity and variability.

Populomics, an Emerging E-Health Response to Comtemporary Healthcare Realities

Michael Christopher Gibbons
Johns Hopkins Urban Health Institute, USA

INTRODUCTION

There is increasing interest in the role of technology in health care and public health. Several researchers have predicted a coming revolution in health care similar to the transformations that occurred in the finance and retail industries upon the widespread adoption of computer technology (Abrams, 2006; Crane & Raymond, 2003; Gibbons, 2005, 2006). In biomedical research, computer technology has catalyzed the emergence of whole new disciplines such as regulomics, proteomics, phenomics, and pharmacogenetics (Gibbons, 2005). Telemedicine, e-prescribing, Electronic Medical Records, and computerized physician order entry systems are emerging as important tools to improve the quality of health care even though the broad health care transformation historically envisioned has not yet occurred (Crane & Raymond, 2003).

Several important factors, however, are suggesting a need for significant changes in traditional health care and medical research systems. These include (1) the growing proportion of people living over the age of 65; (2) the increasing prevalence of chronic diseases; (3) increasing global urbanization; and (4) the increasing recognition of disparities (inequalities) in health and health care (IOM Committee on Quality of Healthcare in America, 2001). In the United States over the last century, many acute and communicable diseases have either vanished or become much less prevalent. Over the same time period, there has been a rise in the prevalence of chronic conditions and diseases. To date, approximately 60% of UK citizens and 50% of US citizens report having at least one chronic disease (IOM Committee on Quality of Healthcare in America, 2001; National Health Service, 2004). Cardiovascular disease is now the number one cause of mortality in developed nations. Although chronic diseases usually result in symptoms and/or death in the later decades of life, the origins of these diseases can often be traced to the first decades of life. To complicate matters further, the elderly often have more than one chronic condition at the same time.

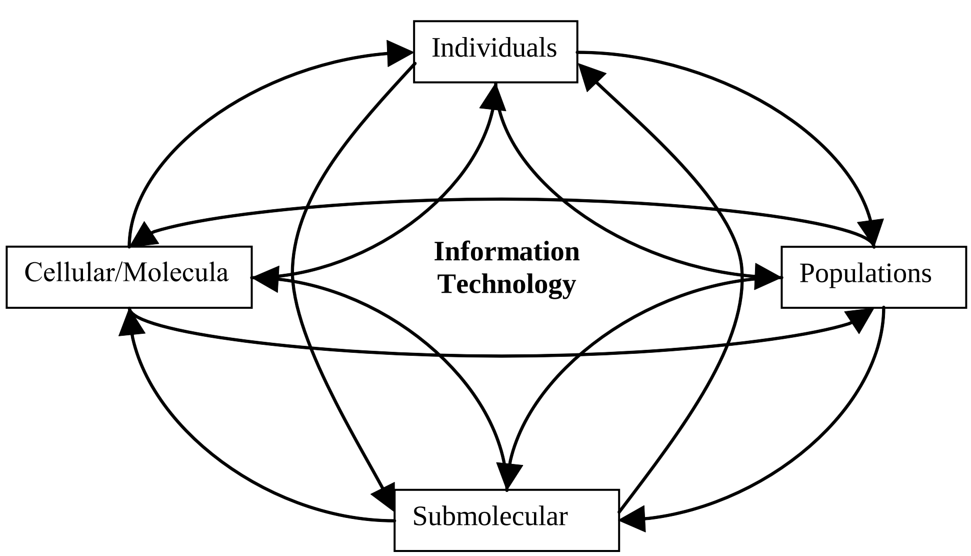
The rapid increase in urbanization that is occurring worldwide is increasingly causing many people living in the inner cities to experience an urban health penalty. This is due to the concentration of economic decline, job loss, and major health problems often found in urban centers (Andrulis, 1997). In addition, significant racial and ethnic disparities (inequalities) are often found in the urban environment. These disparities appear resistant to interventions and policies designed to reduce or eliminate them (Acheson D, 1998; Macintyre, 1997; Smedley, Stith & Nelson, 2003) Increasing evidence suggests that disparities arise as a result of complex interactions among socioeconomic factors; behavior, biologic, and environmental factors; and disease that are related to race and ethnicity (Smedley et al., 2003; Haynes & Smedley, 1999; Faber & Krieg, 2002; Amick, Levine, Tarlov & Walsh, 1995; Evans, Barer & Marmor, 1994). As such, multifaceted approaches that extend beyond the current medical model are needed to improve health status (Andrulis, 1997), particularly in the urban environment.

In general, the major health care systems of the world function as acute disease treatment systems. Chronic diseases, however, occur and progress over long periods of time, often without obvious personal health impact. They then slowly lead to progressive declines in health, punctuated by intermittent episodes of acute illness. Effectively managing the acute episodes of a chronic disease often does little to the natural progression of the disease or the prevalence of the disease in the population. Effectively preventing illness and promoting health among patients with chronic diseases will undoubtedly require a technological infrastructure that can enable the real-time monitoring of physiologic, behavioral, and environmental data simultaneously among large numbers of patients. All this will need to occur prior to the need for hospitalization and also during sometimes lengthy periods of time between acute episodes of illness.

To improve health in the community setting, many kinds of novel health technologies will likely be required. For instance, new spatial analytic, epidemiologic, and computational methodologic innovations will be required to support clinical decision-making by turning vast amounts of information and data into usable knowledge. Computer technology will be essential to the conduct of complex, large-scale clinical trials to objectively implement interventions and assess outcomes. To be successful, this technology infrastructure must also facilitate the transfer of culturally and linguistically appropriate health information to patients. The ability to tailor information to individual patient needs will be critical to adequately supporting patient health choices and self-care. Finally, computer-based innovations such as nanotechnology, distributed processing, RFID, and advanced sensor technology, will facilitate the ability to detect changes in the physical status of things, enable intelligent interactions between devices, and enable devices to make independent decisions based on collected data, devoid of prejudice, bias, or other "human error." "Smart devices" in health care include smart phones that can automatically communicate patient data to physicians, "intelligent" homes that can detect if a senior citizen has fallen and then notify Emergency Medical Services (EMS), "smart" vehicles that can minimize motor vehicle accident related trauma, and wearable "smart" clothing that can monitor and wirelessly transmit environmental and physiologic data to providers. Scientists are even attempting to develop intelligent ovens and online refrigerators that can be controlled through phones or the Internet (International Telecommunications Union, 2005). Such technology applications in health care will enable "smarter," behaviorally oriented, patient-centered, and therefore potentially more effective prevention, public health, and clinical treatment strategies.

Achieving this vision, however, will require a paradigm shift in the conceptualization and provision of health care services and the conduct of medical research. It will require less dependence on reductionism while increasingly relying on a more comprehensive multifaceted approach to scientific inquiry. The term Populomics has been used to articulate a new integrated, transdisciplinary approach to health research and practice (Abrams, 2006; Gibbons, 2005). The term is derived from the synthesis of the disciplines that comprise the field; namely, the population sciences, medicine, and informatics. More specifically, Populomics is defined as an emerging discipline focused on population level, transdisciplinary, integrative disease/risk characterization, interdiction and mitigation that relies heavily on innovations in computer and information technologies. Populomics seeks to characterize the interplay of pathways and mechanisms that work across levels of existence to impact the health of individuals and populations. Populomics focuses on the interrelatedness of "environments" and their complex codependence on each other, which in the aggregate either protect health or enable disease. On the other hand, a populomics orientation will also facilitate the development of transdisciplinary strategies and inter-

Figure 1. Populomics "web of disease/health causation"



ventions to develop socioenvironmental and cultural context-sensitive treatments and regulatory policies that will influence the "web of causation" and ultimately improve health outcomes (see Figure 1.0).

In summary, future health strategies and interventions, whether medical, behavioral, or public health, will increasingly need to rely on current and emerging Information and Computer Technologies (ICT) not only to understand complex health problems but also to help design, implement, and evaluate the solutions and health outcomes. ICTs can be seen as enablers of the discipline of populomics in the same way they catalyzed the development of other scientific and medical disciplines (e.g., regulomics, phenomics, etc.). In so doing, the promise of preventing disease and illness among individuals, promoting health among populations, and eliminating health disparities (health inequalities) may one day become a reality.

REFERENCES

Abrams, D.B. (2006). Applying transdisciplinary research strategies to understanding and eliminating health disparities. *Health Educ Behav, 33,* 515–531.

Acheson, D. (1998). *Independant inquiry into inequalities in health.* The Stationery Office.

Amick, B.C., Levine, S., Tarlov, A.R., & Walsh, D.C. (1995). *Society and health*. New York: Oxford University Press.

Andrulis, D.P. (1997). The urban health penalty: New dimensions and directions in inner city health care. In *Inner city health care*. Philadelphia: American College of Physicians.

Crane, R., & Raymond, B. (2003). Fulfilling the potential of clinical information systems. *The Permanente Journal, 7,* 62–67.

Evans, R.G., Barer, M.L., & Marmor, T.R. (1994). *Why are some people healthy and others not? The determinants of population health.* New York: Aldine De Gruyter.

Faber, D.R., & Krieg, E.J. (2002). Unequal exposure to ecological hazards: Environmental injustices in the Commonwealth of Massachusetts. *Environ Health Perspect, 110*(Suppl 2), 277–288.

Gibbons, M.C. (2005). A historical overview of health disparities and the potential of eHealth solutions. *J Med Internet Res, 7,* e50.

Gibbons, M.C. (2006). Health inequalities and emerging themes in compunetics. *Stud Health Technol Inform, 121,* 62–69.

Haynes, M., & Smedley, B.D. (1999). *The unequal burden of cancer*. Washington, DC: National Academy Press.

International Telecommunications Union. (2005). *The Internet of things* (Rep. No. 377-05). Geneva, Switzerland: The International Telecommunication Union.

IOM Committee on Quality of Healthcare in America. (2001). *Crossing the quality chasm: A new health system for the 21st century*. Washington, DC: National Academy Press.

Macintyre, S. (1997). The Black Report and beyond: What are the issues? *Soc Sci Med, 44,* 723–745.

National Health Service. (2004). *Chronic disease management: A compendium of information*. London: UK Department of Health.

Smedley, B.D., Stith, A.Y., & Nelson, A.R. (2003). *Unequal treatment: Confronting racial and ethnic disparities in healthcare*. Washington, DC: National Academies Press.

KEY TERMS

Biomedical Research: Research at the confluence of biology and medicine.

Chronic Disease: Diseases of long duration with generally slow progression.

Information and Computer Technologies (ICT): Computers and telecommunications technologies.

Public Health: Health care issues that impact all citizens.

Regulomics: Economics of regulations.

Web of Human Interactions: Includes a consideration of disparate areas such as genetic, biologic, behavioral, psychological, and socioenvironmental issues.

Prevalence of Bullwhip Effect in Hospitals

Kannan Sethuraman
Melbourne Business School, Australia

Devanath Tirupati
Indian Institute of Management, India

INTRODUCTION

Growing concern about health care costs in developed countries in the past two decades has spawned extensive research on the industry and, in particular, hospital management. The literature on the subject is extensive and covers a wide range of themes that include application of operational research-based tools, redesign/re-engineering of delivery systems and clinical procedures, and so forth. However, in our case studies, we found that hospitals in Australia are in early stages of adoption of emerging paradigms in operations management (e.g., coordination, lean principles) that have significant potential for enhanced performance. The purpose of this chapter is to describe one of the consequences of poor coordination—*bullwhip effect* in hospitals—and its impact on performance and to discuss initiatives to mitigate and manage it.

In their widely cited papers, Lee, Padmanabhan, and Whang (1997a, 1997b) define *bullwhip effect* as "the phenomenon where orders to the supplier tend to have larger variance than sales to the buyer (i.e., demand distortion), and the distortion propagates upstream in an amplified form (i.e., variance amplification)." The bullwhip phenomenon has been observed by many firms across a number of diverse industries.

Lee, et al. (1997b) state the impact of increased volatility as, "Distorted information from one end of the supply chain to the other can lead to excessive inventory investment, poor customer service, lost revenues, ineffective transportation, and missed production schedules." Although there is a growing body of research on managing the bullwhip effect in manufacturing-based supply chains (Baganha & Cohen, 1998; Chen, Drezner, Ryan & Simchi-Levi, 2000; Chen, Ryan & Simchi-Levi, 1997; Metter, 1997), little research exists on its presence in service chains, and we are unaware of any reported research on this subject. In this chapter, we present several examples of distorted information in hospitals resulting in variability amplification and causing inefficiencies similar to the bullwhip effect. We highlight the underlying causes for this phenomenon and propose actions that can mitigate the detrimental impact of this distortion.

BACKGROUND

The analysis of the bullwhip effect was originally initiated by Forrester (1958, 1961), who was also the first to simulate the demand amplification characteristic through the "beer distribution game." Taking a behavioral perspective, Sterman (1989) and Senge and Sterman (1992) attribute the amplified order variability observed in the game to players' systematic irrational behaviors or misconceptions about inventory and demand information and lack of "systems thinking" by management.

Unlike Sterman (1989), Lee, Padmanabhan, and Whang (1997a) demonstrated that even rational decision-making by individuals can result in the bullwhip effect. They identified four main factors as contributors and offered prescriptions based on four complementary themes: (a) data transparency and information sharing; (b) coordination and synchronization (e.g., demand and supply planning); (c) efficiencies at the operational level; and (d) inventory management for the entire supply chain.

A conspicuous characteristic of the literature on bullwhip effect is its focus on manufactured goods and the role of inventory. Comprehensive approaches for managing supply chain do involve service elements (e.g., interface between retailers and customers). However, consistent with its focus on manufactured products, the emphasis in bullwhip effect literature is on inventory-related service measures such as product availability, fill rate, and so forth.

Relatively very little or no research has been reported to highlight the existence of this phenomenon in service context. There are at least three reasons why research in this area has been relatively sparse. First, simultaneity in production and consumption of services often precludes use of inventory strategies, and thus, many of the results of bullwhip effect literature do not extend readily to services. Second, higher levels of demand variability in services due to seasonality and uncertainty are well recognized, and there exists extensive literature on demand and supply strategies for managing this variability. This literature, which predates emergence of supply chain management, includes research related to price and related incentives for demand and capacity adjustments for supply management. In contrast, service chains typically do not have inventory but only have backlogs that are managed indirectly through service capacity adjustments. Further, in many services, production and consumption happen simultaneously; the service chain comprises only one level, and the issue of amplitude propagation up the chain becomes irrelevant. However, it may be noted that materials and supplies required for the provision of services may exhibit variability amplification, and related results may have direct application.

The aforementioned suggests that the bullwhip effect phenomenon may not be as pronounced and relevant for managing services. However, in our study of institutions in the health care industry in Australia, we have observed several examples of variability amplification similar to the bullwhip effect. The purpose of this chapter is to highlight the existence of bullwhip effect in hospital operations, identify its causes, and prescribe ways to mitigate its detrimental impact.

RESEARCH METHODS

A case-study-based approach was employed to conduct the research. Our findings are based on in-depth case studies of three hospitals—two public and one private not-for-profit. Data were gathered through interviews, observations, and archival sources. Interviews with doctors, nurses, administrators, and other hospital staff were conducted in person. We used an informal, minimally structured, nondirective interview approach to minimize the influence of our assumptions.

Saint Mary's Hospital's Service Value Chain for Elective Surgeries[2]

In this section, we describe briefly the service value chain for elective surgeries at St. Mary's Hospital (SM), one of the hospitals in our case studies, to provide the necessary background for our presentation. It will become clear that other services such as emergencies and outpatients will only make the system more complex and accentuate the bullwhip effect.

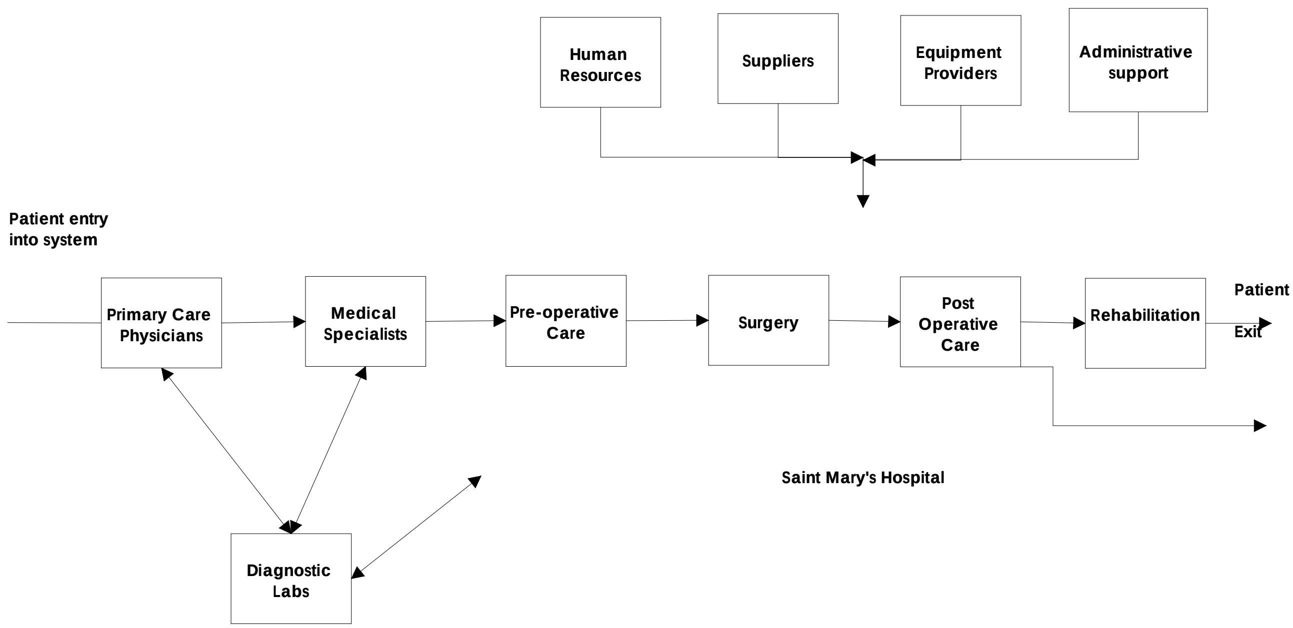
SM, a private not-for-profit hospital, essentially provides facilities comprising of operating rooms (ORs) with state-of-the-art equipment, wards for in-patients, rehabilitation unit for recovery of acute cases beyond normal postoperative care, and related ancillary services. The wards are organized by specialty and considered to be departments, as is the OR management. Hospital administration is decentralized with autonomy for the departments at the operational level. Thus, departments are responsible for nurse staffing and allocation of other resources. Regulatory requirements specifying threshold levels for nurse-patient ratios impose constraints in both ward and OR schedules.

SM does not employ any specialists/surgeons but has only a small group of physicians for overall supervision of the wards. Instead, it has arrangements with a large group of consultant surgeons who hold privileges to avail of the hospital facilities for treatment of their patients. SM derives its revenues from the fees charged for the use of its facilities. In most cases, these are borne by patients' medical insurance providers.

Figure 1 provides a schematic of SM's service value chain. The process begins with a consultation with the patient's primary care physician. If necessary, the patient will visit a specialist for further assessment. This stage may involve several visits to the specialist and diagnostic tests. The patient is added to the doctor's surgery list, if needed. Typically, specialist doctors have privileges at several hospitals in the area, and the choice of hospital is determined by several factors that include cost, patient preference, complexity, wait, facilities, and other services provided by the hospital. Following the choice of hospital, the patient is scheduled for surgery by being assigned a slot in one of the doctor's theater sessions at the hospital.

Surgery represents the third stage of the value chain, and the hospital is responsible for providing required support services, including pre- and postoperative care. Insufficient capacity and/or poor management

Figure 1. Service value chain for surgery patients at St. Mary Hospital



can result in delays for the patient and underutilization of resources. Postoperative care in the hospital and in subsequent rehabilitation, if necessary, represent the fourth and fifth stages before the patient is discharged and exits the hospital system.

Planning and Scheduling at Saint Mary's Hospital

Planning and scheduling at SM is similar to that in other hospitals in Australia with annual theater plans forming the basis for the hospital's activities. The plan involves assignment of theater sessions to surgeons and essentially defines the demand for the hospital's services. A half-day slot is considered the basic unit for this purpose, and thus, each theater has a capacity of 10 sessions per week, based on a five-day week. Surgeons expect hospitals to provide complete flexibility in organizing their sessions, and hence, assignment of a session commits the hospital for the operation and pre- and postoperative care, and thus determines the workload.

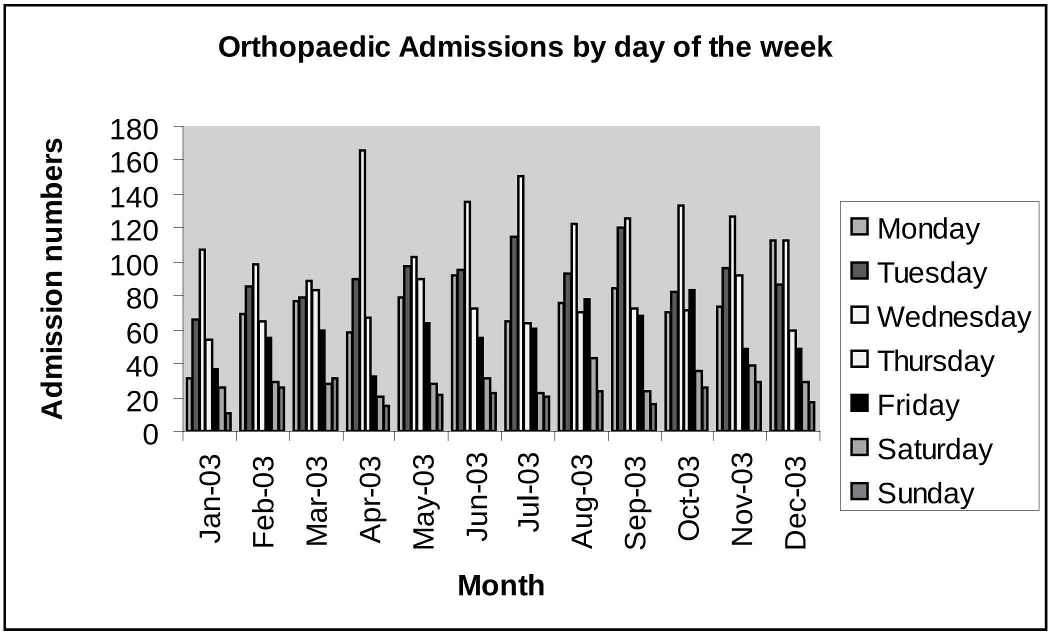
In practice, the annual plan is developed three months in advance, in consultation with the surgeons. The plan can be characterized as cyclic with effort to evenly spread out sessions assigned to individual surgeons. Further, to the extent feasible, weekly schedules are adopted. For example, a surgeon with 50 sessions would be assigned one session per week, usually in the same time slot every week. One feature of session plan at SM that merits mention relates to the practice of concentration of sessions assigned to particular specialty. As a result, sessions corresponding to various specialties peak on different days of the week. Such practices are quite widespread at other hospitals in Australia.

The annual session plan forms the basis for nurse and technician staffing in the hospital. The labor cost of nurses and technicians is perhaps the most significant controllable operating cost at SM and directly impacts the hospital's financial viability. Staffing plan for nurses in the OR theaters is prepared for each calendar month, two weeks in advance, to conform to regulatory requirements. The initial plan is developed on the basis of the session plan. The nurse schedule is revised and frozen only a day in advance, at which time the requirements are assessed based on surgery lists provided to the OR department. At this stage, due to the limited availability of permanent, full-time staff, the nurse scheduler relies on part-time agency nurses to meet the gap in requirements. Nurse staffing in wards exhibit similar trends, and schedules are frozen only a day in advance.

Observations and Comments

Our interviews indicated that in most specialties, there is no pattern in the demand for elective surgeries. This is

Figure 2.



in contrast to illnesses such as flu and hay fever, which exhibit pronounced seasonality. Thus, daily demand for elective surgeries (i.e., customer demand) can be assumed to be uniform displaying no seasonality (day of the week, month, etc.) and low variability. Figure 2 describes the pattern of admissions for orthopedic surgeries at SM by day of the week. It is based on data for one year and excludes the holiday period (mid-December to mid-January). It shows mild seasonality by month (admissions range from a low of 332 to a peak of 512 with a mean of 470) and high variability by day of the week. Excluding the weekends, the average daily admissions range from a low of 58 on Fridays to a peak of 123 on Wednesdays. The pattern is similar for other specialties, except for the location of the peak. Taken in conjunction with our premise that end-customer demand is fairly uniform, this strongly suggests the presence of a phenomenon similar to the bullwhip effect. In the remainder of this section, we discuss briefly its impact on performance—operational and financial.

First, increased variability in surgeries performed directly impacts the demand for postoperative services in the wards. Consequently, on peak days, shortage of beds makes the wards the bottleneck, thereby restricting the number of surgeries, reducing theater utilization and throughput.

Second, the demand for nursing in theaters and wards is affected adversely by higher variability. With SM's flexible strategy of meeting the nursing demand with a mix of full-time, part-time, and agency staff, this results in larger use of casual and agency staff. Higher wage rates for these categories increases operating costs, leading to lower profits.

Third, SM's reliance on flexible staff affects the hospital's operating performance in a more subtle but significant manner. Higher levels of temporary staff result in frequent changes in the composition of support staff assigned to each theater, thereby inhibiting development of cohesive support teams. This factor was mentioned repeatedly by consultant surgeons in our interviews as the primary cause of inefficiencies in the theater, requiring more operating time and reducing theater throughput. Furthermore, this factor influences the doctor's choice of hospital, prompting use of SM for more complex cases requiring its facilities. For simpler cases, the doctors tend to choose more efficient hospitals. As the more complex cases typically involve higher costs, this behavior has an adverse effect on SM's finances. Under the Australian system, the fees payable (to the hospital and doctor) are based on predetermined rates for each class and do not depend on the case complexity within each class.

To summarize, increased variability in the demand for elective surgeries at St. Mary's Hospital results in lower efficiencies, higher operating costs, and lower revenues, leading to lower profits.

CAUSES AND STRATEGIES TO COPE WITH THE BULLWHIP EFFECT

Next, we describe the causes of the bullwhip effect in hospitals and suggest appropriate mechanisms to cope with them.

Demand Forecasting and Updating

In services such as hospitals, where patient presence is an integral part of the service delivery process, using inventory to buffer the demand-supply gap is not possible, and service targets are met through capacity. Thus, in contrast to the manufacturing sector where forecasts are used to plan for inventories, their role in hospitals is to help plan for capacity. With facilities fixed in the short term, capacity is managed through provision of human resources.

At SM, nurse staffing in wards is based on targeted occupancy levels rather than on any forecast. Similarly, nurse staffing in OR theaters is based on annual session plan. Currently, these are developed two weeks in advance to conform to regulatory requirements. The actual daily schedules are determined one or two days in advance based on surgery lists for the ORs and patient status in the wards. Typically, these differ substantially from the plans, and the gaps are met primarily through temporary part-time and agency staff. Currently, 25% of the hospital's nursing requirements is met by agency staff with consequent increase in operating costs.

Batching

Some of the practices at SM are motivated by batching considerations and happen at two levels. First, consultants prefer half-day OR sessions (either morning or afternoon), leading to batches that correspond to fixed time slots. The batch size is variable, depending on the specialty and complexity of the case mix. Thus, consultants' preferences for half-day sessions may be interpreted as batching resulting from implicit setup costs of consultants associated with their visit to the hospital. Second, SM's session plans with concentration by specialty can be interpreted as batching. This practice may have had some justification in the past when volumes were low and theater management was simplified with economies gained in procurement/hiring of equipment from outside. However, this is unnecessary in the current environment and seems to continue due to organizational inertia, resulting in ward beds becoming a bottleneck and leading to lower throughput and related adverse consequences.

Lack of Information Sharing and Coordination

In this section, we present several examples of lack of coordination and information sharing both in planning and operations resulting in the bullwhip effect.

Operations

At SM, the admissions department is responsible for working with consultants' offices and obtaining the necessary details to ensure that adequate resources are available for patient care. The business office is responsible for completing administrative formalities and financial checks prior to the commencement of surgery. Appointments for elective surgeries are made several weeks in advance, and the doctor's office is expected to provide the surgery list to the hospital several days in advance. However, this list is considered tentative (changes account for less than 10%) and not shared with other departments until it is frozen, usually one or two days in advance. This delay in sharing information results in several adverse consequences. First, instead of scheduling patient arrivals in a staggered manner based on surgery sequence, all patients are given a common reporting time and processed in the order in which they arrive. Second, because of the delay in getting the information, the business office is not in a position to complete the checks in advance. Together, these factors delay patient preparation and start of surgery. Third, the lack of information increases the need for flexibility in providing the required nursing and technician resources, thereby increasing costs and leading to other inefficiencies in the theater operations.

Planning

We also observed lack of coordination at the planning stage. First, daily activity schedule and shift times in various departments of the hospital are not synchronized, causing avoidable delays and inefficiencies. For example, the change of shift for nurses in the wards coincides with the surgery times in the theater. Nurses, instead of preparing patients for surgery, are busy with shift change activities and causing patient to wait and delaying start of surgeries. Delayed start induces late arrival among consultants and encourages them to schedule shorter lists (relative to theater capacity) leading to lower throughput.

Second, a coordinated approach for session planning that explicitly recognizes the impact of theater assignments on wards and takes into account bed capacity offers significant potential for increasing the theater output and improving capacity utilization. An integrated approach for session planning is likely to result in a more uniform distribution of theater sessions assigned to one specialty, and implementation of such plans would require buy-in from the consultants. Such an approach can explicitly factor cost and revenue contributions of individual consultants in session planning and help in developing and maintaining a consultant pool consistent with the hospital's strategic objectives, both business and clinical.

Methods for Coping with the Bullwhip Effect

In this section, we outline strategies for mitigating the bullwhip effect (see Table 1 for a summary):

1. **Reducing uncertainty.** If patient information is centrally managed and all parties are provided with the latest data in a timely manner, it will help eliminate duplication of effort and ensure data consistency across all groups. Dynamic updating of this information will greatly help reduce uncertainty.
2. **Variability reduction.** Localized practices that cause increased variability should be eliminated through improved coordination. Use of good forecasting methods will also assist the reduction of variability.
3. **Reducing batching.** Avoiding concentration of theater session by specialty will reduce the batch size. Batching effect can also be reduced by considering shorter time slots for surgeries (e.g., two hours instead of four).
4. **Developing strategic partnerships.** Developing strategic partnerships with key stakeholders is critical. For example, alliances with medical specialists can change the way in which information is shared and help foster coordination.
5. **Realignment of incentives.** The evaluation and reward systems should be modified to stress cooperation across departments to facilitate integrated planning based on system-level objectives.

Table 1. Causes, contributing factors and countermeasures for the bullwhip effect in health care sector

Causes	Contributing Factors	Countermeasures
1. Demand Forecasting and updating	• Lack of willingness to forecast and share based on partial information • Lack of effective use of forecasts in planning and updating of plans	• Have a common database that will minimize inconsistencies across stages • Continual updating of data
2. Batching	• High transportation costs • High setup cost for surgeon's visit to the hospital • Economies in the management of operating theaters	• Modify the measurement and evaluation systems to facilitate the use of systemwide objectives • Reduce the setup cost or transaction cost using lean practices
3. Lack of information sharing and coordination	• Poor incentives for sharing information • No culture of sharing information • Inconsistent information technology platforms across links in the service value chain • Lack of willingness to share until complete confirmation of information	• Alignment of incentives to encourage data sharing • Developing strategic partnership with different links in the service value chain

SUMMARY AND FUTURE RESEARCH

In this chapter, we have identified and described a phenomenon in hospitals that is similar to the bullwhip effect observed in the manufacturing sector. The resulting demand distortion and variability amplification likewise causes performance degradation. However, the similarity with the manufacturing sector does not extend much further since the causes and the impact are somewhat different. For example, in hospitals, it results in lower levels of throughput, higher operating costs, and longer patient waits. While we provide suggestions for improvement based on the case studies, more research is needed to further investigate the root causes. Further, there is opportunity to develop planning tools and systems that recognize, counter, and minimize the impact of the bullwhip effect. Based on our case studies, we conjecture that a similar phenomenon may be present in other segments of the health care sector. Further research in the area can provide more insights into the phenomenon and has the potential for order-of-magnitude performance improvements similar to those realized during the past two decades in the manufactured goods industries.

REFERENCES

Baganha, M.P., & Cohen, M.A. (1998). The stabilizing effect of inventory in supply chain. *Operations Research, 46*, S72–S83.

Chen, F., Drezner, Z., Ryan, J.K., & Simchi-Levi, D. (2000). Quantifying the bullwhip effect: The impact of forecasting, leadtime, and information. *Management Science, 46*, 436–443.

Chen, F.Y., Ryan, J.K., & Simchi-Levi, D. (1997). The impact of exponential smoothing forecasts on the bullwhip effect [working paper]. Northwestern University.

Forrester, J.W. (1958). Industrial dynamics—A major breakthrough for decision making. *Harvard Business Review, 36*(4), 37–66.

Forrester, J.W. (1961). *Industrial dynamics*. Cambridge, Massachusetts: MIT Press.

Lee, H.L., Padmanabhan, V., & Whang, S. (1997a). Information distortion in a supply chain: The bullwhip effect. *Management Science, 43*(4), 546–58.

Lee, H.L., Padmanabhan, V., & Whang, S. (1997b). The paralyzing curse of the bullwhip effect in a supply chain. *Sloan Management Review*, Spring, 93–102.

Metter, R. (1997). Quantifying the bullwhip effect in supply chains. *Journal of Operations Management, 15*, 89–100.

Senge, P.M., & Sterman, J.D. (1992). Systems thinking and organizational learning: Acting locally and thinking globally in the organization of the future. *European Journal of Operational Research, 59*(3), 137–145.

Sterman, J.D. (1989). Modeling managerial behaviour: Misperceptions of feedback in a dynamic decision making experiment. *Management, Science, 35*(3), 321–339.

KEY TERMS

Batch Size/Batching: Refers to the practice of processing several units together to take advantage of economies of scale. In the context of hospitals, the surgeons' preferences for long theater sessions with several surgeries scheduled in one session is an example of batching that is motivated by the time and effort involved in the surgeons' visits to the hospital.

Bottleneck: The limiting resource(s) that constrains the performance of a system. Typically, bottleneck(s) refers to the resource or the constraint that limits the output of the system.

Bullwhip Effect: Represents the tendency for demand variability to increase, often considerably, as you move up the supply chain (from retailer, to distributor, to factory, to raw material suppliers, etc.).

Coordination: Refers to managing the various entities in the value (supply) chain to optimize the performance of the value chain. While synchronization focuses on the individual processes of the system, coordination deals with managing the various entities of the chain. In the hospital service chain context, the surgeons, hospital departments, patients, pharmacists, and so forth are examples of entities whose activities/processes need to be coordinated to improve the chain performance.

Service Value Chain: Represents the chain of activities that generates value to the customer. In this chapter, the hospital service value chain captures the journey that a patient travels through the hospital in achieving the required care.

Synchronous Operations: Refer to management of the entire process together in harmony to achieve the goals of the system. Synchronization involves coordination of all resources, subsystems, and processes so they are in harmony in order to achieve the total system performance, not on local performance measures.

ENDNOTES

[1] This is an abridged and revised version of our earlier publication " Evidence of bullwhip effect in healthcare sector: Causes, consequences, and cures," *International Journal Services and Operations Management*, Vol. 1, No. 4, 2005.

[2] The hospital's name has been disguised to protect its identity.

P

Process Level Benefits of an Electronic Medical Records System

Abirami Radhakrishnan
Morgan State University, USA

Dessa David
Morgan State University, USA

Jigish Zaveri
Morgan State University, USA

INTRODUCTION

The challenges faced by U.S. health care system are vividly explained in the U.S. Government's health information technology plan,

The U.S. health care system faces major challenges. Health care spending and health insurance premiums continue to rise at rates much higher than the rate of inflation. Despite spending over $1.6 trillion on health care, there are still serious concerns about preventable errors, uneven health care quality, and poor communication among doctors, hospitals, and many other health care providers involved in the care of any one person. The Institute of Medicine estimates that between 44,000 and 98,000 Americans die each year from medical errors. Many more die or have permanent disability because of inappropriate treatments, mistreatments, or missed treatments in ambulatory settings. It has been found that as much as $300 billion is spent each year on health care that does not improve patient outcomes – treatment that is unnecessary, inappropriate, inefficient, or ineffective. All these problems – high costs, uncertain value, medical errors, variable quality, administrative inefficiencies, and poor coordination – are closely connected to inadequate use of health care information technology. (U.S. Federal Government Health Information Technology Plan, 2004).

These challenges have resulted in formidable and compelling pressures on the U.S. Federal Government, professional bodies, and accreditation authorities to undertake measures to bring about significant improvements in the quality of health care in both inpatient and outpatients settings. Quality is defined by the Institute of Medicine as having three overlapping domains: safety of patients and practitioners, practice consistent with current domain knowledge, and patient centered care (Kohn, Corrigan, & Donaldson, 2000). In the current environment, one of the important measures undertaken to improve quality of health care is adoption and use of computers and information technology such as electronic medical records system (EMR) by health care providers and medical institutions.

Introduction of EMR has received considerable attention because of its potential effectiveness in (a) implementation of decision support, (b) reducing practice expenses, (c) increasing revenues by improving office efficiency, and (d) making health records more up to date, accessible, legible, and modifiable (Gill, Ewen, & Nsereko, 2001; Hippisley-Cox, Pringle, Carter, Wynn, Hammersley, & Coupland, 2003; Singh, Servoss, Kalsman, Fox, & Singh, 2004).

Even though a few studies in the literature have tried to examine the usefulness of EMR at organizational level (Gill et al., 2001; Hippisley-Cox et al., 2003; Miller & Sim, 2004; Singh et al., 2004), very little research has been done on examining the various components of EMR systems and the process level benefits of these systems. As Barua, Kriebel, and Makhopadyay (1995) point out, by attempting to relate IT implementation directly to output variables at the organizational level, the intermediate processes through which IT impacts are felt are totally ignored. When information technologies such as EMR systems are deployed effectively, they interact with intermediate organizational processes, and deliver value by building unique process level benefits, which finally result in organizational level outcomes. Therefore, there is a need for more granular studies that explicate the underlying linking mechanism (process) between IT implementation and organizational level impacts (Radhakrishnan, Zu, & Grover, 2007).

In this article, we pursue this line of thought and examine the various components of EMR and the potential process level benefits of EMR systems by presenting case summaries. This article contributes to the literature and practice in two ways: First, it provides insights into various modules of EMR system and how they interact with each other to bring out value to users. Second, we present six case summaries (based on qualitative case analysis and content analysis of Web sites) to illustrate the potential process level benefits of EMR systems. The remainder of this article is organized as follows. The next section provides an overview of EMR and its components. In the following section, we present six case summaries to illustrate the potential process level benefits of EMR systems.

OVERVIEW OF EMR

What is EMR?

Electronic medical records system (EMR) is defined as an interorganizational information system that captures the essential components of a patient's medical encounter with the medical provider, including storage and retrieval of subjective, objective patient information, and assessment and plans for patient care (Lenhart, Honess, Covington, & Johnson, 2000). EMR systems facilitate physicians and other health care professionals to:

- Monitor the health status of their patients with electronic medical charts.
- Support health care decisions with evidence-based guidelines.
- Expedite referrals to specialists and other health care providers.
- Computerize ordering of prescription drugs, laboratory tests, and radiology results.
- Store and retrieve patients' medical records from different locations.

Why EMR?

While conducting a field study involving hospitals in countries such as UK and USA, Lederman (2005) found that there are several problems faced by health care professionals while handling patients' medical records. These problems are listed below:

First, in many hospitals, only paper-based records are maintained. Management of paper-based records becomes a significant problem due to the possibilities of losing or misplacing records.

Second, data inconsistency and data integrity are other problems related to paper-based records. Patient files can be changed with no assurance that changes made would be incorporated into a central repository or vice versa. This could lead to a possibility of different, conflicting records being held for the same patient. This increases the probability of incorrect information being accessed.

Third, nonintegrated databases across hospitals pose a significant problem, providing a major impediment to collection, access, and ability to view patient records. The medical staff in each hospital manually enters patients' records into the system. This may cause potential problems like inefficient use of time, loss of productivity, and data entry errors.

EMR can help in overcoming these problems (Bria, 2006; Doyle, 2006; Simon & Simon, 2006). In a survey of 703 practicing family physicians in the U.S., Karsh, Beasley, and Hagenauer (2004) found that the physicians preferred electronic medical records to paper based records as they were up to date, modifiable to meet individual needs, accessible when needed that resulted in better record quality. These studies highlight the value of EMR systems.

Components of EMR

An EMR system has several different modules. Organizations can implement specific modules or all the modules depending upon their needs.

Patient Registration Module (PRM)

This module helps with patient's registration. It helps track and maintain relevant details on the patient (such as name, date of birth, address, contact number, financial class), health insurance plans (insurance company, type of insurance policy, and level of copay), and historical as well as scheduled appointments.

Appointment Scheduling Module (ASM)

This module helps with scheduling appointments. Appointment reservations are scheduled for specialists, general practitioners, and for other resources such as medical labs, emergency rooms, operation theaters, and so forth. It provides flexible appointment schedules where one can request convenient time slots. Patients and support staff can search and view availability of resources within various departments, specialties, and locations.

Medical Billing and Receivables Module (MBRM)

This module helps with "super-bill" (i.e., comprehensive bill) creation and collection of receivables. This module helps track charges associated with diagnosis, medical check-ups, medical procedures, surgical procedures, lab tests, and medical prescriptions. Then super-bills are created by bundling all the charges for a variety of services provided by different departments. The co-pay amounts from patients are recorded. This module retrieves patients' insurance details and sends medical claims to insurance clearance agencies. This module tracks all the accounts receivables from insurance agencies. If payments are not received from insurance agencies within the stipulated period, it rebills the customers for outstanding dues.

Ambulatory Care Module (ACM)

This module maintains patient's medical records. It maintains patients' family and social history, statistics on vital parameters, immunization records, allergies, surgical procedures, physical examination reports, diagnosis results, lab reports, radiology reports, EKG charts, MRI, CT scan reports, medication reports, and any other remarks from other physicians in a central repository. This module permits users to record and retrieve digital photos, transcriptions, drawings, radiology images, MRI, CT scan charts, texts, graphs, and voice. Physicians can have their dictated progress notes transcribed and imported into the EMR, or record handwritten progress notes using a tablet PC.

E-Prescription Module (EPM)

This module generates and tracks electronic prescriptions. Physicians can write prescriptions electronically, which are then transmitted to the associated formularies and retail drug stores. Physicians can even electronically order referrals and laboratory and radiology tests. Users can track test-order status and extent to which a particular prescription is filled by a drug store.

Workflow Management Module (WMM)

This module manages the workflow within a health care organization. This module creates work lists for various professionals such as Specialists, General Practitioners, Nurses, and Lab Staff. This module interacts with the Appointment Scheduling Module and creates work lists for various professionals who are working in a health care organization.

Knowledge Base Module (KBM)

EMR system maintains a repository of clinical references for various disorders, medical complications and tracks information on latest drug dosages and delivery systems. Medical professionals can search this repository for any reference information.

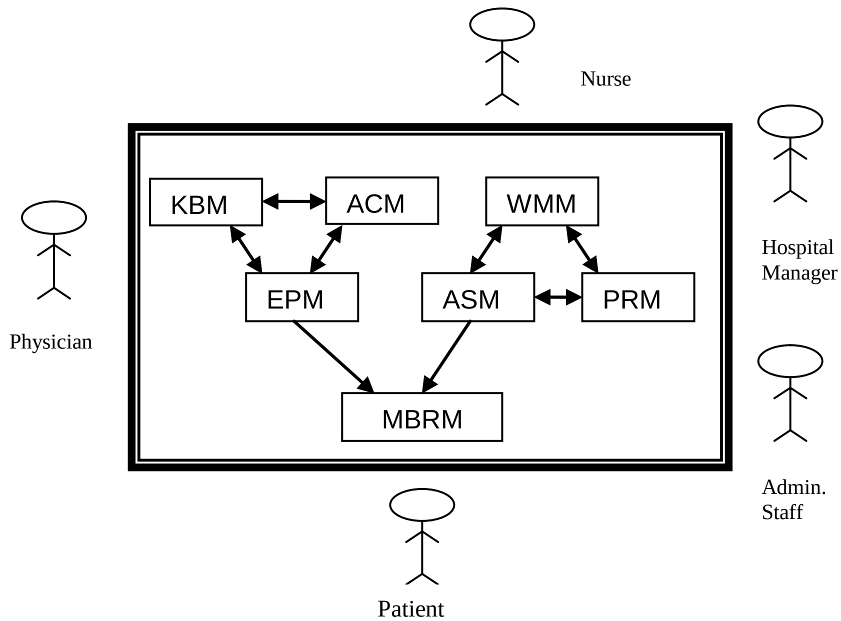
Figure 1 shows the different modules and their interaction. The appointment scheduling module (ASM), patient registration module (PRM) and workflow management module (WMM) frequently interact with each other while knowledge base module (KBM), ambulatory care module (ACM) and e-prescription module (EPM) frequently interact with each other. ASM and EPM feed into medical billing and receivables module (MBRM). As shown in Figure 1, physicians, nurses, hospital managers, administrative staff, and patients use EMR system.

RESEARCH METHODOLOGY

We conducted an exploratory case study on process level benefits of EMR systems. A case study is appropriate for an exploratory investigation like this (Yin, 2002). We employed a case study design using six organizations. We conducted interviews with EMR managers, physician champions, nurses, radiologists, and EMR vendors associated with a few hospitals and physician organizations that had adopted an EMR. We conducted the interviews during the period 2005-06. Using six different organizations allowed comparisons between them for more general research results (Benbasat, Goldstein, & Mead, 1987; Yin, 2002).

We triangulated qualitative data from multiple sources. As mentioned earlier, we conducted telephonic interviews with EMR managers, physician champions, nurses, radiologists, and EMR vendors. We relied

Figure 1. Modules of EMR system and their inter-relationship



heavily on the support of these experts. We used a structured interview guide for our interviews. This interview guide had both closed ended questions and open ended probing questions. Findings from closed ended responses were compared with answers to the open-ended questions. We verified answers by having interviewees review our notes on their responses. We also conducted several face-to-face meetings with representatives from one hospital who were willing to participate in our project.

We also reviewed reports from secondary sources such as trade journals. We also conducted content analysis of Web sites of organizations. We considered data associated with the period 2001-2006. Krippendorf (1980) defines content analysis as a technique wherein a researcher searches for structures and regularities in the content and draws inferences on the basis of these regularities. Here content refers to process level benefits. For consistency, the same researcher did content analysis of the various Web sites.

We conducted cross-case analysis and identified consistent patterns regarding process level benefits of EMR systems.

ANALYSIS AND DISCUSSION OF PROCESS LEVEL BENEFITS

We present six case summaries to highlight the process level benefits of EMR systems.

Case Summaries

Case #1: Oconee Memorial Hospital (http://www.oconeememorial.org/index.cfm?PageID=72)

Oconee Memorial Hospital operates in South Carolina, USA, and is considered progressive in the use of health care IT systems. In late 2003, they got interested in implementing an EMR system known as Practice Management Plus and Horizon Ambulatory Care. It is a part of a larger information network known as Foothills Health Information Network. They initially implemented the system in the radiology section in early 2004 before using it hospital wide. Radiologists used EMR to create and interpret images derived from a variety of modalities: radiography, ultrasound, magnetic resonance imaging (MRI), positron emission tomography (PET), and computed tomography (CT). EMR captures both analog and digital images. EMR also captures metadata concerning who viewed, annotated, printed, and sent specific data and when those activities occurred. This ability of EMR to capture and replay aspects of invisible work resulted in information richness as well as new accountabilities. The search and retrieval processes were enhanced in a digital environment. Now this hospital uses many modules such as appointment scheduling, billing and receivables, patient registration, ambulatory care, and workflow management. Based on the post-implementation evaluation, they have identified the following process level benefits: improved coordination among departments within the hospital, improved access to the patients' medical records from remote locations, improved throughput of diagnosis, treatment and consultation, reduced time taken for delivery of services, improved speed and accuracy of medical insurance claims registration, improved data integrity of patient's medical records, better control of employee work schedule and the institution's overall responsiveness.

Case # 2: Ogden Clinic (http://www.nextgen.com/images/pdfs/Ogden-FCG.pdf)

Ogden Clinic is a multispecialty group practice in Utah, USA. They implemented Next Gen EMR and enterprise practice management systems in 2005. Ogden Clinic began experiencing the following process level improvements: Patient check-in process gets completed in half the time and several reporting tasks associated with the receptionist function have been reduced by half. Using point-of-care documentation templates have allowed Ogden Clinic to eliminate 95% of its transcription use. Scanning and electronic storage of documents have reduced the volume of paper work by 75%. Electronic ordering and results management have been greatly streamlined—with an electronic interface to the laboratory information system, an EKG interface, and e-prescribing capabilities saving valuable physician and support staff time. With health maintenance reminders linked to its disease management programs and a robust set of online patient education materials, Ogden Clinic automatically generates patient recall reminders on a weekly basis, alert physicians when patient interventions are overdue, and provide comprehensive clinical information tailored to each patient's condition. As a result, the organization has recorded an annual increase in revenue of $72,000 attributable to the additional health maintenance visits. Whenever a national drug is recalled, Ogden Clinic generates a list of all its patients on that medication and sends them personalized letters with instructions, providing timely care and saving staff time.

Case # 3: Northwest Diagnostic (www.healthmgttech.com)

Northwest Diagnostic is a family practice clinic in Texas, USA. They implemented the clinical charting, ambulatory care, billing, and scheduling modules of the e-MD EMR system in 2005. Viewing capabilities of EMR system resulted in an up to date availability of data and charts, better data organization and legibility. Support staff reported that they spend less time finding, pulling and filing charts. Physicians reported that they spend less time in locating information. They were able to access and modify the records as and when needed. They reported that they are able to spend more time on assessment and patient care. Coding documentation (different medical procedures have different codes) also improved. Quality of preventive care and disease management improved because of use of electronic alerts on symptoms and allergy lists. Billing procedures also dramatically improved. All these resulted in a saving of a quarter of million dollars in the very first year after implementation.

Case # 4: Cayuga Family Medicine (http://www.healthmgttech.com/archives/1106/1106bridging_gap.htm)

Cayuga Family Medicine is a New York-based low volume practice that focuses on pediatrics and obstetrics. They implemented the patient registration, appointment scheduling and medical billing modules of EMR system in 2000. Cayuga realized the following process level improvements: reduction in the transcription and billing errors, improvement in the speed and accuracy of medical insurance claims registration, data consistency of the patient's medical records and overall productivity of the physicians and front office staff.

Case # 5: Newport Heart (http://www.alteer.com)

Newport Heart is a California based cardiology unit. It implemented the medical billing, clinical charting, and patient registration modules of the EMR system. They experienced the following benefits after implementation of the EMR system: improvement in cash flows, reduction in accounts receivables by 25%, reduction in the time taken for delivery of services to patients, improved ability to inform patients about their up to date financial account status, accurate patients' records, and reduction in the paper work by 90%.

Case # 6: Public Hospitals and Polyclinics under National Health Care Group and Singapore Health Services, Singapore http://www.zdnetasia.com/news/business/0,39047112,39189253,00.htm)

Seven public hospitals under National Health Care Group and 17 polyclinics under the Singapore Health Services share in-patient discharge summarizes including prescriptions, treatments, and allergies and out-patient records, x ray and lab reports through a centralized platform of EMR system. The hospitals and polyclinics have realized the following benefits: better coordinated care for patients moving across different providers, reduction in transcription errors, better clinical decisions with access to complete and legible clinical histories, cost savings through the avoidance of unnecessary repeat tests and investigations, 24-hour access to real time data for up-to-date results reporting, and improved data security with provision for audit trails.

In summary, we were able to identify several process level benefits from EMR adoption and use such as:

a. Improved coordination between departments in the hospital,
b. Better access to legible medical records from remote locations,
c. Increased throughput for diagnosis, treatment and consultation,
d. Increased speed and accuracy of medical insurance claims registration,
e. Improved data integrity of patient medical records,
f. Better work schedule management,
g. Improved data security and audit trails,
i. Better disease management and preventive care, and
j. Increased costs savings.

CONCLUSION

Electronic Medical Records system (EMR) is an enabling interorganizational information system for overall improvement in health care quality. In this article, various modules of EMR and the relationships between them were presented. There are several modules such as patient registration, appointment scheduling, ambulatory care, e-prescription, knowledge base management, work flow management, billing and receivables management modules. They all work synergistically to increase the value and benefits to all stakeholders.

The process level benefits of EMR systems were then examined by conducting qualitative case analysis of six organizations that had implemented EMR systems and content analysis of their Web sites. Assessment of benefits (at the process level) helps to explicate the underlying linking mechanism between EMR implementation and organizational level impacts. We were able to identify several process level benefits from EMR adoption.

While a few studies in the literature tried to examine the usefulness of EMR at organizational level, very little research has been done to examine the various components of EMR and process level benefits of EMR systems. We have set a stage for examining process level benefits of EMR systems to organizations. Researchers should further this line of research by conducting a longitudinal study involving several health care organizations and examine the process level benefits of EMR systems. Researchers could use theoretical lenses of Transaction Cost Economics, Resource Based View, and Digital Options theory to explain the process level benefits of EMR systems.

REFERENCES

Anonymous. (2007). Newport heart case study. Retrieved February 14, 2008, from http://www.alteer.com/features+benefits/case_studies/index.php

Barua, A., Kriebel, C.H., & Mukhopadyay, T. (1995). Information technologies and business value: An analytical and empirical investigation. *Information Systems Research, 6*(1), 3-23.

Benbasat, I., Goldstein, D.K., & Mead, M. (1987). The case research strategy in studies of information systems. *MIS Quarterly, 11*(3), 369-386.

Bria, W. (2006). Applied medical informatics for the chest physician: Information you can use! *Chest, 129*, 446- 452.

Davis, M. (Ed.). (2007). Foothills health information network. Retrieved February 14, 2008, from http://www.oconeememorial.org/index.cfm?PageID=72

Doyle, M. (2006). Promoting standardized nursing language using an electronic medical record system. *AORN Journal, 83*, 1335-1347.

First Consulting Group. (2007). *Measurable benefits achieved early in EMR implementation – and growing quickly.* Retrieved February 14, 2008, from http://www.nextgen.com/images/pdfs/Ogden-FCG.pdf

Gill, J.M., Ewen, E., & Nsereko, M. (2001). Impact of electronic medical record on quality of care in primary care office. *Deleware Medical Journal, 73*(5), 187-194.

Hippisley-Cox, J., Pringle, M., Carter, R., Wynn, A., Hammersley, V., & Coupland, C. (2003). The electronic patient record in primary care – regression or progression? A cross sectional study. *British Medical Journal, 326*, 1439-1443.

Karsh, B., Beasley, J.W., & Hagenauer, M.E. (2004). Are electronic medical records associated with improved perceptions of quality of medical records, working conditions or quality of working life? *Behavior and Information Technology, 23*(5), 327-335.

Kohn, L.T., Corrigan, J., & Donaldson, M.S. (Eds.). (2000). *To err is human: Building a safer health system* (1st ed.). Washington DC: National Academy Press.

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Beverly Hills, CA: Sage Publications.

Lai, E.N. (2004). Tech supplements. Retrieved February 14, 2008, from http://www.zdnetasia.com/news/business/0,39047112,39189253,00.htmLederman, R. (2005). Managing hospital databases: Can large hospitals really protect patient data? *Health Informatics Journal, 11*, 201-210.

Lenhart, J.G., Honess, K., Covington, D., & Johnson, K.E. (2000). An analysis of trends, perceptions, and use patterns of electronic medical records among U.S. family practice residency Ppograms. *Family Medicine, 32*(2), 109-114.

Loehr, J. (2007). *Bridging the gap*. Retrieved February 14, 2008, from http://www.healthmgttech.com/archives/1106/1106bridging_gap.htm

McBride, M. (Ed.). (2007). Northwest diagnostic case study. Retrieved February 14, 2008, from www.healthmgttech.com

Miller, R.H., & Sim, I. (2004). Physicians' use of electronic medical records: Barriers and solutions. *Health Affairs, 23*(2), 116-126.

Radhakrishnan, A., Zu, X., & Grover, V. (in press). A process-oriented perspective on differential business value creation by information technology: An empirical investigation. *Omega: The International Journal of Management Science*.

Simon, S.J., & Simon, S.J. (2006). An examination of the financial feasibility of electronic medical records (EMRs): A case study of tangible and intangible benefits. *International Journal of Electronic Healthcare, 2*, 185-197.

Singh, R., Servoss, T., Kalsman, M., Fox, C., & Singh, G. (2004). Estimating impacts on safety caused by the introduction of electronic medical records in primary care. *Informatics in Primary Care, 12*, 235-241.

U.S. Federal Government Health Information Technology Plan. (2004). Retrieved February 14, 2008, from http://www.whitehouse.gov/infocus/technology/economic_policy200404/chap3.html

Whitten, J.L., & Bentley, L.D. (2007). *Systems analysis and design methods*. New York: Mc Graw Publications.

Yin, R.K. (2002). *Case study research, design and method* (3rd ed.). Newbury Park: Sage Publications.

KEY TERMS

Electronic Medical Records System: An interorganizational information system that captures the essential components of a patient's medical encounter with the medical provider, including storage and retrieval of subjective, objective patient information, assessment, and plans for patient care (Lenhart et al., 2000).

EMR Process Level Benefit: A benefit measured at the process level after 2 years of adoption and use of Electronic Medical Records System in medical institutions (adapted from Radhakrishnan et al., 2007).

Health Care Information System: An arrangement of information technology, people, data, and processes that interact to gather, process, store, and disseminate health care information (adapted from Whitten & Bentley, 2007).

Production Management in the Elderly Care Services

Nora Ekroos
Helsinki University of Technology, Finland

Erkki Vauramo
Helsinki University of Technology, Finland

INTRODUCTION

In western countries, the so-called demographic time bomb, that is, the ageing of the baby-boom generation, has become one of the most challenging issues. Although it has become almost clichéd in health care planning, its effects are being felt rather acutely in reality. The situation in Finland, as in many other western countries, is compounded by the fact that as demand for *elderly care* is increasing, the service systems are suffering from severe labor and tax funding shortages. In fact, population in Finland is aging faster than any other OECD country (Antolin, Oxley, & Suyker, 2001). Elderly care centers have difficulties in hiring qualified professional staff. Nursing staff are also burdened by heavy workloads. The situation will worsen by time as the number of elderly people in our population increases further, leading to increased strain on health care resources. The present service structure is not going to be able to respond to this demand. Yet health care funding, which depends on public financing, will decrease as the number of taxpayers declines due to the aging of our workforce. "Elderly dependence ratio," a key demographic indicator, will approximately double over the next two decades (Eurostat, 2005).

We have approached this problem described above from a *production management* point of view. Some ideas from this concept can be implemented to improve elderly care, such as work-in- progress. This article reports on the findings of a case study using this approach in the field of elderly care in Helsinki City, 2004. A previous version of this article has appeared in the HCTM 2005 Conference proceedings.

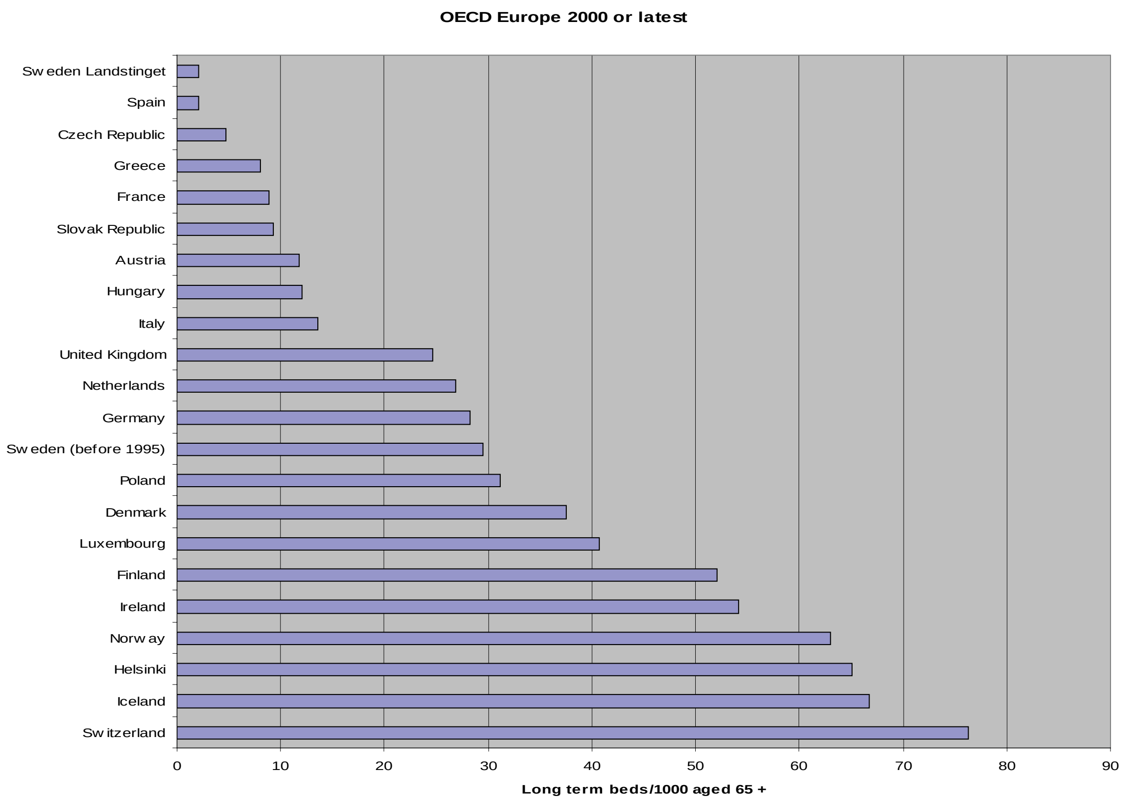
BACKGROUND

In the Finnish care system, the municipalities are responsible for providing health and social care services for all inhabitants including the elderly. This care is financed by local municipal taxation, although this can be supplemented by additional funding from central government.

The amount of long-term beds is relatively high in Finland (see Figure 1). In 2002, 5 % of the over 75s lived in service housing, 5 % in nursing homes, and 3 % were in-patients at health center long-term wards. In Sweden, the health care reform has accomplished a radical reduction in institutional care. Most of the Swedish elderly (96%) live in a single room with their own kitchen and toilet.

A shift towards the deinstitution of the elderly can be observed (Coleman, 1995; Eng, Pedulla, Eleazer, McCann, & Fox, 1997). A study performed in Italy (Tibaldi et al., 2004) suggests that severely demented elderly patients at home can benefit from the same level of care provided by a hospital ward. Stable support from family members of demented subjects can delay their admission to nursing home. Maintaining independence remains central to the quality of life of elderly persons with disabling chronic conditions (Ball, Perkins, Whittington, Hollingsworth, King, & Combs, 2004). A study performed in Australia (Allen, Glasziou, & Del Mar, 1999) reported that bed rest is not only used in the management of patients who are not mobile, but is also prescribed as a treatment for a large number of medical conditions. The researchers (Allen et al., 1999) reviewed available long-term care studies and concluded that bed rest does not help to treat any illness, but on the contrary actually worsens the

Figure 1. Long-term beds/1000 aged 65+ in Europe 2000 or later (OECD, 2000)



condition of the patient. Those who stay in the hospital for a very long time are usually those that consume the largest amount of hospital resources (Marshall, McClean, & Millard, 2004). In addition to the expense of long-term institutional care, it is also an inherently dependent way of living.

PRODUCTION MANAGEMENT IN HEALTH CARE

Health care has been perceived as different from any other type of service. The present requirement to improve efficiency is putting pressure on finding new approaches to service delivery. This has been given increased priority as a result of the increasing pressure on elderly care services in particular.

In recent years there has been increased interest in applying *production management* thinking to health care (Lillrank, Kujala, Kämäräinen, & Kronström, 2003; Thompson, Wolf, & Spear, 2003). The lessons learned from the industrial environment have guided innovations in health care service processes. With this in mind, the Department of Industrial Engineering and Management at Helsinki University of Technology has compiled a health care research team designed to transfer industrial engineering principles to the domain of health care.

The work-in-progress (WIP) measure in manufacturing has been introduced to health care through the patient-in-process concept (PIP) (Kujala et al., 2004). The focus in studying PIP is on patient episodes. In industry, process time has been divided into value-adding and nonvalue-adding components, but for health care the categorization needs to be more sophisticated. The proposal (Kujala et al., 2004) is to divide the duration of patient episode into three major components: diagnostic and care time, administrative time, and waiting time. The distinction between time categories is based on the expected change in a patient's medical condition, information management, the types of services provided and resource consumption. Cutting off nonvalue-adding time can lead to a remarkable improvement in a patient's condition and consequently provide savings for municipalities, other employers, and insurance companies.

Next we will be present how the production management approach and the work-in-progress concept were implemented to elderly care service production in our case study.

CASE HELSINKI

In Helsinki City, there are over 550, 000 inhabitants. The proportion of people over age 65 is 13.7 %, while 6.5 % of the population is over 75. There are several elderly care service providers: Helsinki City Health Department, Helsinki City Social Service Department, public and private foundations for elderly care homes, private enterprises, and others. The degree of institutionalization in elderly care is high (see Figure 1). The elderly care expenditure is also significant, €4260 per capita for those over 75, which is about twice as much as in the other cities/municipalities in Finland. This cost disparity can be partly, but note entirely explained by the higher employee and estate costs in the metropolis area.

Helsinki City has produced several analyses and evaluations on elderly care, the main conclusions of which are that some of the elderly are not currently located in the right category of care facility. Relatively healthy elderly people are laying in beds in hospital wards, although their quality of life would be higher in senior residential housing.

Taking cognisance of the challenges related to the optimization of a multi-organization service network, the City of Helsinki and the Department of Industrial Engineering and Management agreed upon a research project to analyze the processes and problems in 24 hour elderly care. This study focused on the processes between various service providers; in other words, the service system was examined from the perspective of the entire care path. The results of which will now be presented.

Research Objectives

The main objective was to analyze how the organizational structure and the service provision processes influence on the cost and quality of elderly care using industrial process analysis methods. The steps were as follows:

1. To describe the 24 hour service structure and operational environment;
2. To analyze the “selecting the next step facility” procedure after an acute care period;
3. To model patient flow by processes in the elderly care service system; and
4. To understand the mechanism driving extra costs incurred by this service system.

Research Methods

The elderly care processes were identified using the available data. The process perspective is a new approach in order to understand the care pathways and the supply-demand relationships of different elderly care services.

The data allowed us to do perform a number of quantitative analyses. The objective of which was to construct a process map to quantify the system structure and to identify possible bottlenecks or other problems in care pathways. The qualitative data was collected through semi-structured interviews of key persons.

Available Basic Data and Materials

Data for process analysis and reference materials were mainly gathered from databases of the Health Department and the Social Service Department of the City of Helsinki. The national registers of STAKES (The National Research and Development Centre for Welfare and Health) were used as reference materials. The financial statements of the Social Services Department and Health Department for 2003 were also used. The key person interviews were performed with a focus on the current service structure, collaboration between organizations, decision making, and follow-up procedures.

Findings

Service Structure and Operational Environment

The Health Department as well as the Social Service Department are responsible for the provision of 24 hour elderly care. Helsinki produces most of the care services on its own but also buys some nursing home beds and service housing beds from private provid-

Table 1. Elderly care providers in Helsinki

Provider	Amount of Locations
Health Department	*4*
Social Service Department	*20*
City owned Helsinki Nursing Home Foundation	*6*
Several Third sector of private care home	*89*
Helsinki Service Housing Ltd (org. for city owned residences)	*16*

ers and foundations (see Figure 2). The complicated administrative governance structure creates challenges in order to offer frictionless care pathways. There are five types of organization that provide 24 hour elderly care serves (see Table 1).

There are four different levels of care: acute hospital care, institutional care, service housing, and home care (see Figure 2). Long-term ward, nursing home and 24 hour service housing are all included in the 24 hour elderly care. The production expense of nursing homes is €93 per day, 24 hour service housing is €55 per day, while at €145 per day the long-term ward is the most expensive of these care locations. In long-term ward the care is mostly based on bed rest. For the elderly patient, the nursing home and service housing are the most independent places to live in the 24 hour care environment.

Selecting the Next Step Facility

After a period of acute care, the aim is to disband the patient as swiftly as possible. The selection of the patient's next facility is made by a committee consisting of representatives from open and institutional care. This decision making process is based on the patient's care needs as measured by the care index system. In principle the location of elderly patients will be based on the careful estimation of working order.

In the current system there are several problems associated with the effective disbandment of elderly patients. For example, there is no *rehabilitation* location that caters especially for the elderly. The only possible solution is to locate the not-yet-rehabilitated patients in institutional care. Despite the potential for the *rehabilitation* of elderly patients, all are placed in institutional care or service housing with the same protocols.

Care Pathways in Network

The process analysis (see Figure 3) illustrates the real patient flow and explains how bedridden patients are born. The data analysis confirmed that the service structure also interacts with the subsequent stage of facility location. As a result of over-capacity in the long-term hospital, as well as undersized service housing and nursing home facilities, elderly patients who so not require hospital care are located in the long-term ward. The occupation rate of which is mostly 100% or over (extra beds organized) in all 24 hour service levels. As has been established, the maximal capacity utilization does not always offer the most optimal cost effectiveness.

After a period in the long-term ward, it is difficult to get elderly patients up from their beds because their general condition will be weaker as a result of a prolonged period of bed rest. Some of the elderly have to stay in the hospital for many years, although for many of them bed rest constitutes a nonvalue-adding time. Furthermore, their quality of life would be better in a noninstitutional environment. Integrated care pathways do not function because several providers are involved and the service structure does not respond to the demand. This institutional oriented service structure leads to institutional care too early.

The difference between the required number of beds and the beds in use in the long-term ward is illustrated in Figure 4. Approximately 600 long-term ward beds can be closed while simultaneously increasing rehabilitation, service housing, and nursing home beds. If the service structure would response to demand, the resources required in the long-term ward would decrease. This care pathway cross-section analysis revealed one of the bottlenecks in the service system and explains partly the long-term ward periods.

Figure 2. The service structure and the division of service production between social service department, health department and other providers in Helsinki

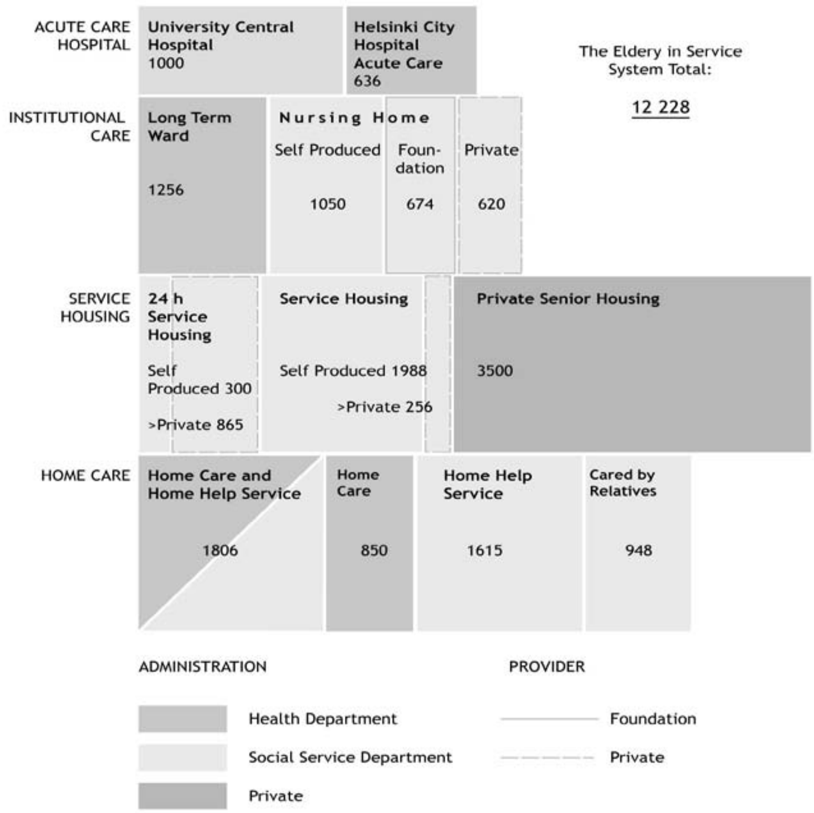


Figure 3. This phenomenon inflicts the long care periods in institutions in Helsinki; 09/2004

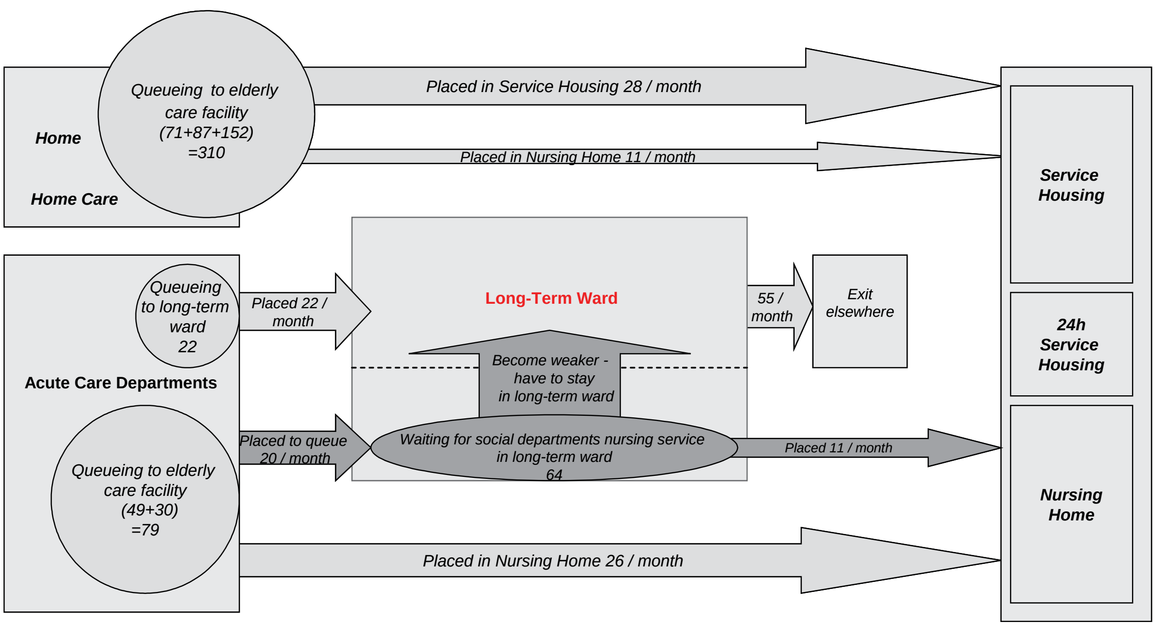
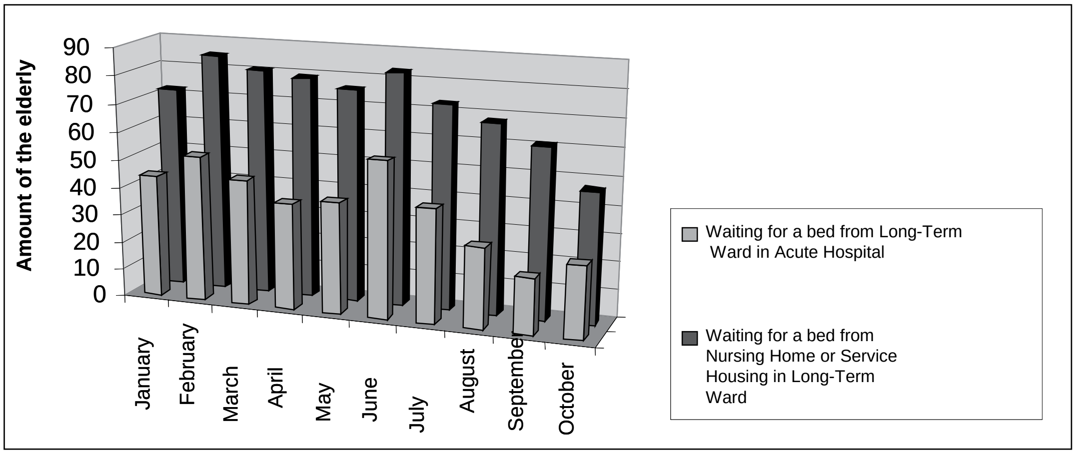


Figure 4. Differences between the beds needed and beds offered in the long-term ward in Helsinki, 2004



Processes Between Services

The study showed that the amount of possible care pathways, as indicated in Figure 2, is over 30 (between different providers and services). As an example, the Helsinki Nursing Home Foundation is owned by the city but is treated as a private provider creating an unnecessary border between services. The analyses of interface functionality were performed and the queues between providers were found out. Inoperative interfaces inflict bottlenecks in the process between services.

All administrative organizations in this case study produce their own budgets and follow-up reports. The entire care costs are not available, making it difficult to understand the cost factors in the service system from the perspective of the entire care path. Communication between organizations should be transparent in order to provide services more efficiency and offer frictionless care path to the elderly. Every interface should be under critical scope and with a systematic follow-up system.

CONCLUSION

It was speculated by the city that the "metropolis problem" was the principle cause of the significant elderly care expenditure. Data has showed that the problems are rooted in both the service system, nonfunctioning processes between the institutions as well as the lack of systematic process management. For example, timely rehabilitation of aged patients following an acute hospital admission is important in order to avoid several years of institutional care. The superfluous cost of elderly care in Helsinki is caused partially by the service structure, which due to the lack of rehabilitation places elderly people in institutional care often years too early.

The findings explain the reasons why the elderly are not always located in the correct facilities in Helsinki. The elderly care process can be categorized as the patient-in-process (PIP) in health care environment. The term "not-rehabilitated-in-process" (NRIP) as a new category can be used to analyze the elderly care processes and the caused costs. Patients that are located in care facilities of unwarranted cost intensity due to unrealized rehabilitation potential suffer from nonvalue-adding waiting time.

The possibility of rehabilitation in the prescribed time during and/or after an acute hospital visit is very important in order to avoid many years of institutional care. The extra costs associated with elderly care in institutional environments are caused often by the elderly people who are steered to institutional care sometimes even years too early as a result of the lack of rehabilitation services. Spending time in institutions does not add value to those elderly patients who still have the potential to rehabilitate. The patient's medical condition does not usually improve with bed rest. The diagnostic and care periods in acute care are completed but the care process is not finished. The administrative time might also be too long in the care process if the service structure does not respond to the demand. The problem is not caused because of a shortage of resources but as a result of the service structure and

Table 2. Care costs in different service levels in Helsinki

	Long Term Ward		Nursing Home		24h Service Housing	
Care Period	Patients +75	Care Cost (million €)	Patients +75	Care Cost (million €)	Patients +75	Care Cost (million €)
0-30 days	479	1,53	172	0,35	56	0,10
30-59 days	131	0,84	93	0,43	40	0,13
60-89 days	74	0,79	90	0,68	36	0,20
90-179 days	116	2,26	260	3,25	102	0,97
0,5-1 years	120	4,71	327	9,09	183	3,66
1-2 years	241	18,69	416	22,8	245	10,32
2-10 years	499	**109,56**	692	98,47	510	60,28
over 10 years	27	15,22	47	23,9	33	12,34
Total	**1687**	**152,92**	**2097**	**158,78**	**1205**	**87,95**

the "steering process" (which reflects the type of care for each elderly patient). As Table 2 illustrates, the 2-10 year care periods in the long-term ward are the most expensive ones. Therefore, the present service structure in Helsinki causes the extra costs because of the nonvalue adding long-term ward periods.

The costs of an unrehabilitated person can be considered analogous to the cost of work-in-progress in production management. Although the acute hospital has treated the acute illness, the patient is not fully rehabilitated. The importance of well organized rehabilitation is one of the key issues in order to avoid institutionalization too early.

FUTURE TRENDS

A process chart is a very illustrative way to analyze elderly care pathways in order to understand the relationship between supply and demand in service systems and also to identify bottlenecks in elderly care processes. Problems in organizational structure and process governance are also easily detected.

Not-rehabilitated-in-process (NRIP) is not only expensive, but also inflicts extra suffering on elderly patients. Service structures should respond to the demand, and follow-up procedures should be reviewed from a process perspective. The present unintegrated follow-up system does not support the frictionless care pathway. It is impossible to control the processes and optimize the resource usage without proper data. Resources should be re-organized and avoid the gratuitous institutional care that is expensive and leads to a dependent way of living.

In the future, as indicated in the study, more economical and user friendly care pathways can be produced. There will be more elderly people to take care of with less money from the tax payers. The importance of delivering well-timed 24 hour elderly care for the right person will increase in order to ensure that everyone in need of 24 hour care will receive it. Organizing the service structure and steering the elderly care processes in the most effective way will be a very important issue to improve the quality of care for the elderly, with optimal use of resources. It is possible to offer better quality of life to the elderly with fewer resources. From our experience, we believe the application of industrial process analyses to health care will be used in the future.

REFERENCES

Allen, C., Glasziou, P., & Del Mar, C. (1999). Bed rest: A potentially harmful treatment needing more careful evaluation. *The Lancet, 354*(9186), 1229-1233.

Antolin, P., Oxley, H., & Suyker, W. (2001). How will ageing affect Finland? Organisation for economic co-operation and development (Economics Department Working Papers, No. 295).

Ball, M.M, Perkins, M.M., Whittington, F.J., Hollingsworth, C., King , S.V., & Combs, B.L. (2004). Independence in assisted living. *Journal of Aging Studies, 18*, 467-483.

Coleman, B.J. (1995). European models of long-term care in the home and community. *International Journal of Health Services, 25*(3), 455-474.

Eng, C., Pedulla, J., Eleazer, G.P., McCann, R., & Fox, N. (1997). Program of all-inclusive care for the elderly (PACE): An innovative model of integrated geriatric care and financing. *Journal of American Geriatrics Society, 45*(2), 223-232.

Eurostat. (2005). Population projections 2004-2050, EU25 Population rises until 2025, then falls, Ref: 48/2005.

Kujala, J., & Peltokorpi, A. (2004, December 3-4). Application of time based management in the improvement of patient process efficiency. In *First Conference of the POMS College of Service Operations: Operations Management in Services: Theory and Practice*. Columbia University, New York.

Lillrank, P., Kujala, J., Kämäräinen, V., & Kronström, V. (2003, September 7-9). Patient in process- a new approach to managing patient processes in health care. In *the Hospital of the Future -3rd International Conference on The Management of Health care & Medical Technology,* Warwick, UK.

Marshall, A.H., McClean S.I., & Millard P.H. (2004). Addressing bed costs for the elderly: A new methodology for modelling patient outcomes and length of stay. *Health Care Management Science, 7*(1), 27-33.

OECD Health Data. (2000). OECD.

Thompson, D.N., Wolf, G.A., & Spear, S.J. (2003, November). Driving improvement in patient care: Lessons from Toyota. *Journal of Nursing Administration, 33*(11), 585-95.

Tibaldi, V., Fabris, F., Molaschi, M., Aimonio, N., Ponzetto, M., Maero, B., et al. (2004). Home care for demented subjects: New models of care and home-care allowance. Archives of Gerontology and Geriatrics, 9(Suppl. 1), 155-162.

KEY TERMS

Acute Care: Short-term care in ward.

Care Path: Illustrates the patient flow through the different care services also over organizational boundaries.

Institutional Care: Long-term care in ward or in nursing home.

Not-Rehabilitated-in-Process (NRIP): The diagnostic and care periods in acute care are completed but the care process is not finished, for example, lack of rehabilitation.

Service Levels: In the elderly care environment, there are four different levels of care, including acute care, institutional care, service housing, and home care.

Programming Body Sensor Networks

Talles Marcelo Gonçalves de Andrade Barbosa
Catholic University of Goiás, Brazil
University of Brasília, Brazil

Hervaldo Sampaio Carvalho
University of Brasília, Brazil

Iwens Gervásio Sene Jr.
Catholic University of Goiás, Brazil
University of Brasília, Brazil

Francisco A. de O. Nascimento
University of Brasília, Brazil

João Luiz Azevedo de Carvalho
University of Southern California, USA

Adson Ferreira da Rocha
University of Brasília, Brazil

INTRODUCTION

The miniaturization and cost reduction of microelectronic devices have been leading to the development of new technologies. Wireless sensor networks (WSNs) are one example of these new technologies. A WSN is a distributed system that is composed of autonomous units with sensing capabilities (sensor nodes), interconnected by wireless communication. WSNs have been successfully applied in the monitoring of the human body. This development led to a new concept: Body Sensor Networks (BSNs).

According to Yang (2006, pp. 5), BSNs have specific requisites, when compared to WSNs. Among these requisites, we highlight the following:

- High level of security, in order to guarantee the confidentiality of information;
- Biocompatibility and biodegradability;
- Higher sensitivity to data loss, and the consequent need for mechanisms to ensure a minimum Quality of Service (QoS);
- Need for context awareness, as the physiological variations are strongly related to the changes in the context in which the user is in;
- Low number of sensor nodes, which should, however, be usually more precise than sensors for other applications of WSNs;
- Requirement of nodes with the capability of running multiple tasks.

Pervasive monitoring demands great adaptation capability from the BSN. Moreover, cases in which the decision made by the system can be different from the decision that would be made by a healthcare professional are frequent. Therefore, besides intelligent algorithms that allow autonomous operation, BSNs need mechanisms that allow changes in their behavior in order to become a clinically useful tool. According to Baldus, Klabunde, and Müsch (2004), "the BSN has to work automatically, but has also to be always under explicit control of any clinician."

Thus, specific models that include programmability as a functional requisite are important in a software architecture designed for BSNs. However, the greatest challenge is to allow the modifications to be made not only to the structure of the software, but also in its behavior, without excluding the capability for autonomous operation of the system.

According to Barbosa, Sene, Carvalho, da Rocha, Nascimento, and Camapum (2006), two important concepts related to programmability in BSNs are: (i) deployment-time programmability, and (ii) run-time set-up. The deployment-time programmability refers to the definition of software artifacts and algorithms that are embedded in the sensor node. In BSNs, the inclusion of this functionality requires a programming interface that is suitable for healthcare personnel, as well as intelligent compilers. Intelligent compilers should be capable of handling implicit functional and nonfunctional requisites of a program. As an example, the inclusion of mechanisms and policies for energy saving could be treated by these structures.

The run-time set-up refers to the capability for adjustments in run-time. The BSN should provide interactivity between the healthcare professional (the BSN manager) and the system. As a requisite, sensor nodes need mechanisms that allow a better control of the tasks that are being run. A possible solution is the use of data structures that allow preemptive multitasking.

The goal of this article is to present the current state of the art, regarding programmability in BSNs. Moreover, we want to present potential benefits of a paradigm shift in which healthcare professionals become the actual programmers and maintainers of the BSNs. With that in mind, we briefly present a software architecture that has been developed with the goal of allowing programmability at network and sensor node levels.

BACKGROUND

Many issues related to software for BSNs have not been discussed yet. Among them, we can mention: (i) the development of graphical user interfaces directed to healthcare personnel; (ii) the hardware abstraction layers (HALs); (iii) the standardization of services and information structuring (BSN ontology); and (iv) programmability at sensor and sensor node levels. The solutions to these problems can lead to improvement of the effectiveness of these systems.

Currently, the software used in BSNs has the following characteristics:

- They are composed of proprietary systems built based on a specific architecture (hardware), and designed for handling a single application. They have little or no modularization, and they are not committed to software development methodologies. In general, they do not employ multiprogramming. Sensor nodes are usually viewed just as sources of data, and all processing is performed in a gateway—an element that interconnects the BSN to other systems—or in a Local Processing Unit (LPU), which is an element to where the data are transmitted. Some examples of such systems are presented in Asada, Shaltis, Reiner, Sokwoo, and Hutchinson (2003), Valdastri, Menciassi, Arena, Caccamo, and Dario (2004), Linz, Kallmayer, Aschenbrenner, and Reichl (2006), Kara, Kemaloglu, and Kirbas (2006), and Chakravorty (2006).
- These systems are usually based on a generic, general purpose model. It is usually based on the NesC programming language (Gay, Levis, von Behren, Welsh, Brewer, & Culler, 2003), on the TinyOS Operating system (Hill, 2003), and on a network programming system, the Deluge (Hui & Culler, 2004). All of these systems are free of charge, and they were developed by the University of California, in Berkeley. CodeBlue (Welsh, 2006), WHMS (Jovanov, 2006), and UbiMon (ICL, 2006) are examples of designs that use the TinyOS framework. Other examples are presented in (Bauldus et al., 2004) and in (Farshchi, Nuyujukian, Pesterev, Mody, & Judy, 2006).

Regarding programmability, systems built based on TinyOS have the following limitations when applied to BSNs:

- The NesC programming language imposes a peculiar syntax, based on concepts emerged from software engineering. Without the knowledge of programming logic and the expertise in managing software components, it is virtually impossible for a nonspecialized user to use this system.
- The multiprogramming model used in TinyOS is not interactive enough. It offers little control over the activities (tasks) run by the sensor node, because there is no context switch. Tasks cannot be immediately interrupted or replaced in order to answer a policy established by the application, or to answer to a command issued by the user. Moreover, according to Han, Bhatti, Carlson, Dai, Deng, Rose, Sheth, Shucker, Gruenwald,

and Torgerson (2005), the execution of more complex tasks can be limited by the occurrence of deadlocks.

BSNs must be designed to operate in an autonomous manner. On the other hand, they should offer mechanisms that yield the control to healthcare personnel, since all the functionalities of the network should reflect their judgment and their clinical evaluation. Thus, a great challenge in BSN software design is to increase, in a transparent way, the access to the internal configurations of the network and its sensor nodes without damaging its capability for autonomous operation.

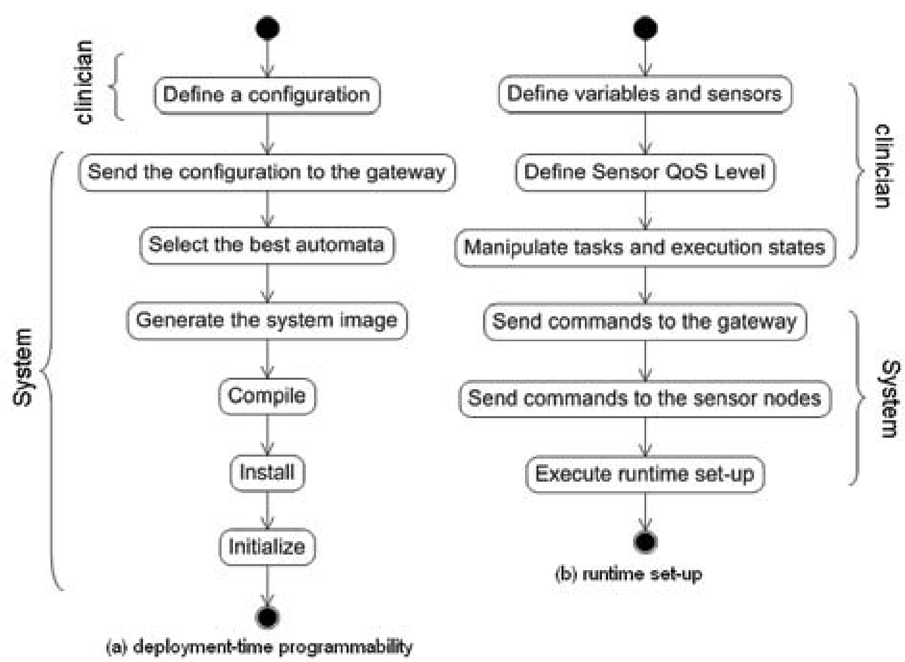
PROGRAMMING BODY SENSOR NETWORKS

Software Architecture for Body-worn Sensor Networks (SOAB) is a system that is being developed at the University of Brasilia to support BSN-based applications (Barbosa et al., 2006). SOAB incorporates the concept of programmability, presented in the introduction of this article, as functionality (service) offered by the BSN. Figure 1 presents a diagram illustrating the activities related to the method for programming BSNs proposed by SOAB. Part of these activities should be run by healthcare professionals (clinicians), and the remaining activities are run by the system itself.

Figure 1a refers to programming at deployment time. These activities are responsible for the definition of mechanisms and policies that are included in the sensor nodes to support a specific application. Figure 1b refers to the run-time set-up process, responsible for the definition and/or adjustments of applications and data sources. The detailed definition of these activities will be presented later, during the description of the layers of the proposed architecture.

SOAB is a modular system, divided into four independent functional layers: (i) a graphical interface, developed especially for healthcare personnel; (ii) middleware for interconnection of the BSN with the Internet; (iii) a server for handling the services related to the services requested by the programmers; and, finally, (iv) an operating system with support for preemptive multitasking, which will be installed in

Figure 1. Activity diagrams to describe the SOAB programming methodology



the sensor nodes. This operating system was named MedOS, and one of its main features is that it tries to increase the lifetime of the sensor nodes by means of proper scheduling of tasks, based on policies that are adapted to biomedical applications. In order to systematize these policies, an automata-based model has been used. Moreover, MedOS allows BSN managers to manipulate the states of the running tasks. Figure 2 presents an overview of SOAB.

In the first layer of SOAB (Figure 2), we have the BWSNET Configuration Tool. This graphical interface is responsible for providing means by which healthcare professionals can describe their algorithms in a way that is more intuitive, less time-consuming, and less error-prone. This interface also allows the possibility of including new sensors that were not initially specified, without the need to recompile the source code. An interface based on Java Reflection (Java Tutorials, 2006) was used to achieve this feature. The BWSNET Configuration Tool has also a simulator that acts as a virtual debugger for estimating the sensor node lifetime and other performance metrics. As a sample of the BWSNET Configuration Tool, Figure 3a shows a deployment-time programming interface for the acquisition of the electrocardiogram (ECG), while Figure 3b shows the run-time set-up interface.

Figure 3b also shows how an application is defined, and how it can be modified. As an example, the application "monitoring user stress" was defined by four variables: blood pressure, ECG, heart rate, and Blood O2. Each variable can also be defined by other variables of data sources, as described by Carvalho, Perillo, Heinzelman, and Murphy (2004). The "ECG" variable, in this case, has been defined by three data sources: ECG 1Lead 100 Hz, ECG 3Leads 500 Hz, and ECG 6Leads 1000 Hz. In order to support these functionalities, the tasks and the software artifacts have also been included at deployment-time (see Figure 3a).

In order to increase the lifetime of the network and, consequently, of the application, the system is responsible for determining how much time each data source should stay active, and how it should operate (the required precision). In order to do so, the system uses information regarding the application in order to execute policies that promote, for instance, energy saving. These policies act on the scheduling of the running tasks and on the adjustment of the bandwidth, and of the transmission power of the communication interface of the sensor nodes (Sensor QoS Level). By using the run-time set-up interface, the healthcare professional can alter the configuration that is usually chosen by the system, excluding and/or adding new data sources and variables, and can manipulate the Sensor QoS Level.

Figure 2. SOAB's four layer system

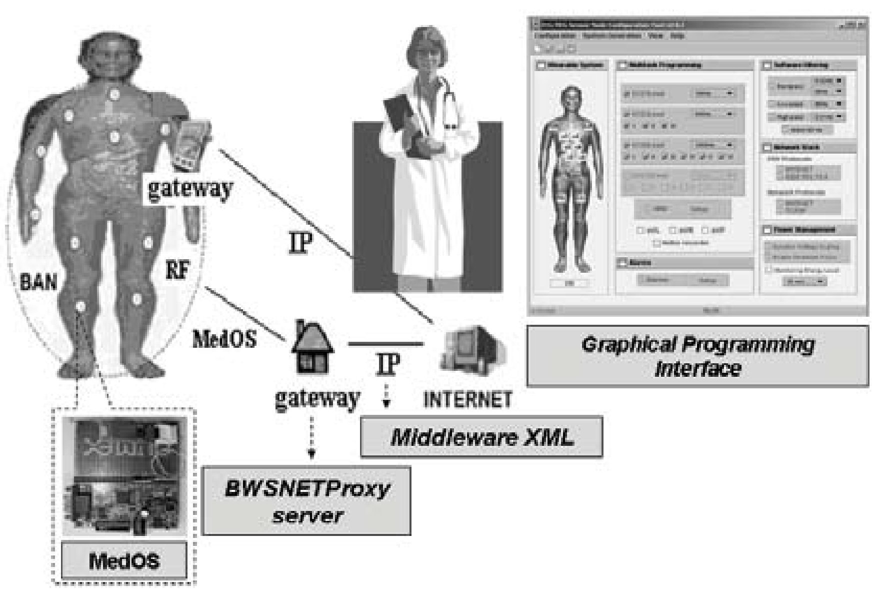
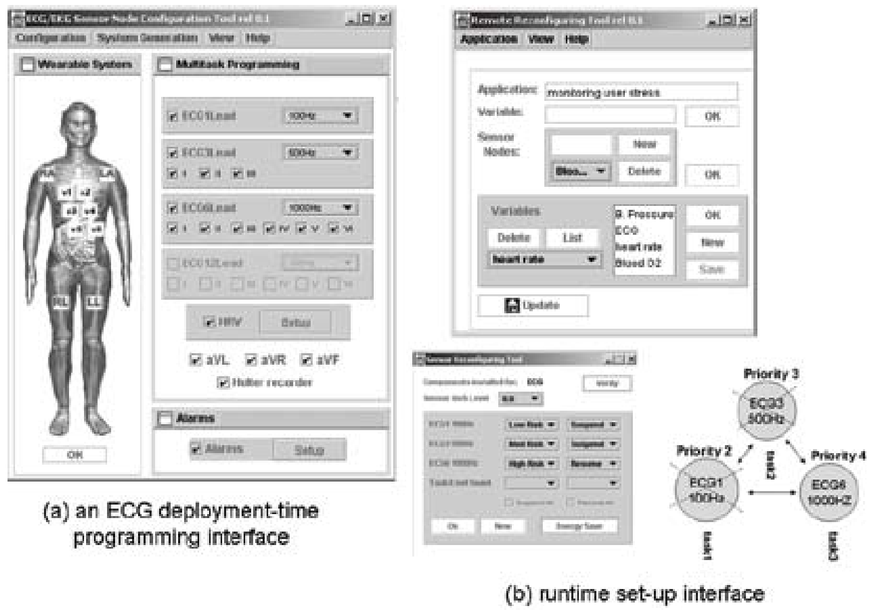


Figure 3. Examples of SOAB graphical interfaces



The user can also change the current values attributed to the priorities associated with the tasks, and also suspend and restart tasks.

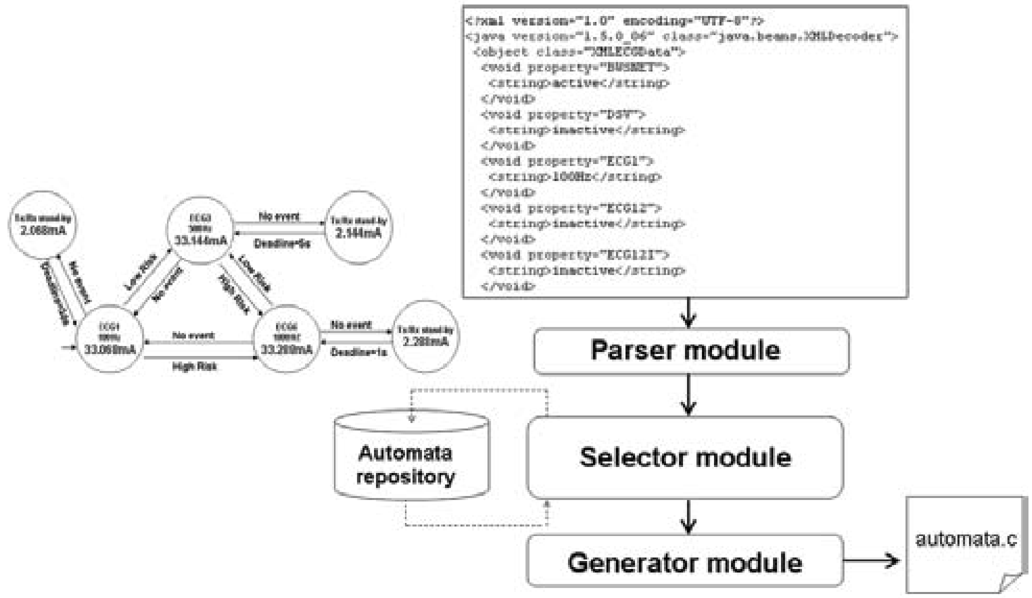
In order to provide support for the programming and reconfiguration of the sensor nodes from the Internet, a group of remote procedure calls (RPCs) has been implemented in compliance with the recommendations of the W3C Web Services (http://www.w3.org/TR/2004/NOTE-ws-arch-20040211/; http://www.w3.org/TR/wsdl20/). These RPCs represent the XML middleware layer in Figure 2.

In the third layer, we have the *BWSNET Proxy*. This software entity has been projected to operate in the gateway devices. It is responsible for the translation of the requisitions of the SOAP XML format (http://www.w3.org/TR/2003/REC-soap12-part1-20030624/) to commands that are understandable to the underlying layer. Besides, it has the responsibility of creating the real image of the system from the BSN manager's program and the correspondent automata, the selfadjustable binary code that will be executed by the sensor nodes. Another functionality of the BWSNET Proxy server that is equally important, is that it promotes the integration of the BSN with an external healthcare information system (HIS).

Figure 4 shows the module that is responsible for selecting the best automatum for a given configuration defined at deployment time. The automata generator is an intermediate-level intelligent compiler. It is composed of three modules. The "Parser" module is responsible for extraction relevant information for the "Selector" module. In practice, these informations are related to parameters, such as the objective of the application, the number of tasks, the types of tasks, and the selected alarms, among others. The values of these parameters are used for filling a table of parameters that will be used by the selector module. The "Selector" module chooses the automatum that maximizes the potential of each application, guaranteeing a greater lifetime for the system. In practice, the selector module is a decision tree (Cormem, Leiserson, & Rivest, 1997, pp. 173) that is responsible for the choice of the best automatum from a set of possible automata that are deposited in an automata repository. This choice is made based on the table of values that is made available by the "Parser" module. The definition of this three (selection algorithm) is based on the medical knowledge that influences the values of each parameter in the table.

After the choosing the best automatum, the "Generator" module can generate the "automato.c" file (see Figure 4), which will represent the program that will be embedded into the sensor node. For this task, a description of the necessary code libraries and necessary system calls should be used. This description is stored in the automata repository, along with the functional description of each automatum. During the automata

Figure 4 . The automata generator module



selection, this configuration file is used by the "Generator" module for assembling the "automato.c" file.

In the last layer of SOAB, there is an application-oriented operating system called MedOS. The main objective of MedOS is to provide support for the behavioral adjustment of the sensor-node at run-time, changing the priority values that can be associated with the tasks provided by the sensor-node. Figure 5 shows the main artifact of MedOS and their dependencies. The functionalities are represented by tasks created together using FreeRTOS libraries (Barry, 2006).

Besides the preemptive multitasking functionality, MedOS implements a set of device drivers, a set of library functions that are used by the automata generator module, and a command interpreter. The command interpreter allows programmers to alter the scheduling of tasks that is currently adopted by the system. This functionality may or may not be included during the generation of the image of the system. If this functionality is disabled, artifacts related to the library "interpreter.h" (Figure 5) will not be included. Thus, only artifacts related to the "policy.h" file (which represent the behavior of the automatum that has been chosen by the application) will be included in the "automata.h" file, which represents the image of the system that will be installed in the sensor nodes.

In addition to the code libraries of FreeRTOS, the MSPGCC code libraries (http://mspgcc.sourceforge.net/) have also been used for the implementation of MedOS.

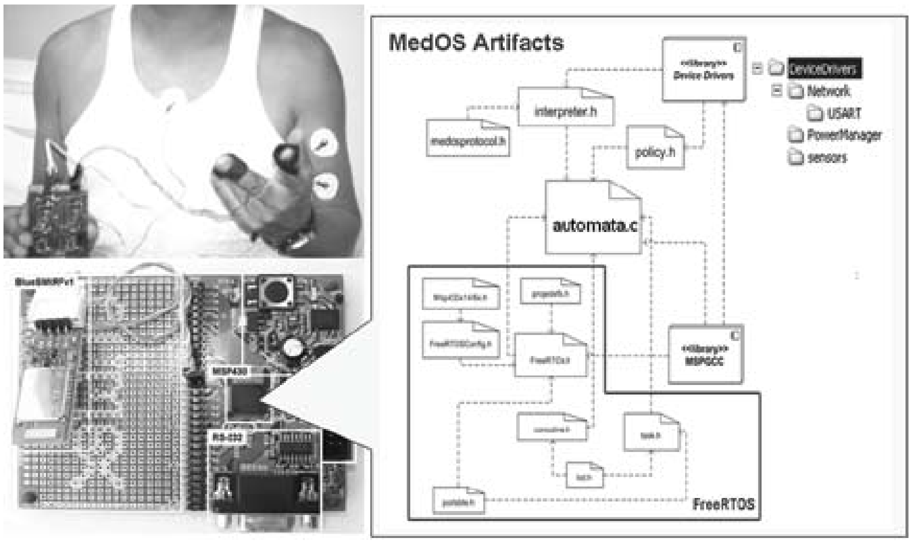
SOAB has been tested in the hardware platform presented in Figure 5. The sensor-nodes were built using the Olimex MSP-P149 kit (http://www.olimex.com/dev/msp-p149.html), with a bluetooth radio BlueSRMIFv1 (http://www.sparkfun.com/commerce/product_info.php?products_id=582). This system allows the monitoring of electrocardiographic and electromyographic signals, as well as skin temperature, arterial pressure, and galvanic skin resistance.

FUTURE TRENDS

The SOAB includes three main innovative features to the BSN research area: (i) a graphical interface for programming and reconfiguration of the sensor nodes; (ii) an application of the concept of preemptive multitasking to BSNs; and (iii) an intelligent compiler (which is also an automata generator). The authors believe that these three features will be pursued by future efforts in this area. Some other potential future trends may be highlighted:

- *The study of man-machine interfaces, and its repercussions on programming and reconfiguration models of BSNs*. A graphical user interface is

Figure 5. MedOS artifacts and sensor node architecture



an attractive and effective alternative, especially for remote access using the Internet. However, in some situations and environments, such as in an emergency room, other devices (hardware) could be used in a more effective way. Baudus et al. (2004) discuss the use of a infrared set-up pen, which works as a remote control, allowing the physicians and nurses to connect the data sources needed for patient monitoring. These data sources are implemented with sensor nodes that can be interconnected to form a BSN that is suitable for the desired application.

- *Smart compilers* that are capable of embedding, during the programming at deployment time of the sensor nodes, mechanisms and policies that increase the effectiveness of the application should also be targeted by new developments. These compilers must be based on a specialist's knowledge, since the medical expert is the one who should know how to schedule the resources according to each application, and should operate in a transparent way to the user.
- *Smart and adaptive algorithms* for task scheduling that take into account the dynamics of the applications should also be pursued. For instance, the health condition of the patient and the capability of the sensor node and of the network should be taken into account in the formulation of these algorithms. In this context, preemptive multitasking allows the computer system to guarantee for each process a regular "slice" of operating time in a more reliable way. It also allows the system to rapidly deal with important external events, such as incoming data, which might require the immediate attention of one or another process.
- *Stantardization of the software features* and of the process related to the development of sytem for BSNs will also be an important issue. In this subject, we can highlight: (i) the development of methodologies for evaluation of these systems regarding their usability and effectiveness; and (ii) the standardization of the interfaces and services as a tool for promoting the independency of hardware and for facilitating the integration of information with other networks and systems. For example, the definition of a BSN Sensor Ontology could establish a hierarchy that expresses the semantics associated to each type of sensor, group or relationship. This approach can increase the effectiveness in the use of BSNs facilitating the gathering of information, which could also be used for programming and reconfiguration of the network.

CONCLUSION

By using a modular software architecture that is designed according to the requirements of possible applications, it is possible to increase the capability for autonomous operation of the BSN, and still offer tools for programming and reconfiguration of these systems that can be used for people with little grasp on programming languages.

This text presented a proposal for a paradigm shift for the programming and reconfiguration of the BSNs. By making more suitable tools available—such as programming interfaces, intelligent compilers, and application-oriented algorithms—the goal is to allow healthcare workers to become the actual programmers and maintainers of this technology. The possibility of developing and/or testing new applications without the need of specific technical knowledge on programming languages and computational models can facilitate the popularization of this technology.

SOAB integrates concepts such as reflection, Web services, automata, intelligent compilers, and preemptive multitasking, with the goal of becoming a solution that is functional, portable, usable, effective, and easy to maintain. Also, this architecture can be extended into four levels, allowing its improvement and the implantation of other systems (services). This solution establishes a concept: application-oriented software architecture.

REFERENCES

Asada, H., Shaltis, P., Reiner, A., Sokwoo, R., & Hutchinson, R. C. (2003, May–June). Mobile monitoring with wearable photoplethysmographic biosensors. *IEEE Engineering in Medicine and Biology Magazine, 22*(3), 28–40.

Baldus, D., Klabunde, K., & Müsch, G. (2004). Reliable set-up of medical body-sensor networks. *Lecture notes in computer science, 2920/2004,* 353–363.

Barbosa, T. M. G. de A., Sene, I. G. Jr., Carvalho, H. S., da Rocha, A. F., Nascimento, F. A. O., & Camapum, J. F. (2006). Application-oriented programming model for sensor networks embedded in the human body. In *Proceedings of the 28th Annual International Conference IEEE Engineering in Medicine and Biology Society (EMBC)* (Vol. 1, pp. 6037–6040). IEEE EMBC Press.

Barry, R. (2006). *FreeRTOS*. Retrieved February 6, 2008 from, http://www.freertos.org/

Carvalho, S. H., Perillo, M., Heinzelman, W., & Murphy, A. (2004, January–February). Middleware to support sensor network applications. *IEEE Network Magazine Special Issue, 18*(1), 6–14.

Chakravorty, R. A. (2006). Programmable service architecture for mobile medical care. In *Proceedings of the Fourth IEEE International Conference on Pervasive Computing and Communications WorkShops (PERCOMW'06)*. IEEE Computer Society Press.

Cormem, T. H., Leiserson, C. E., & Rivest, R. L. (1997). *Algorithms* (1st ed.). Cambrige, Massachusetts: MIT Press.

Farschchi, S., Nuyujukian, H. P., Pesterev, A., Mody, I., & Judy, W. J. A. (2006, July). TinyOS-enabled MICA2-based wireless neural interface. *IEEE Transactions on Biomedical Engineering*, *53*(7), 1416–1424.

Gay, D., Levis, P., von Behren, R., Welsh, M., Brewer, E., & Culler, D. (2003). The nesC language: A holistic approach to network embedded systems. In *Proceedings of the ACM SIGPLAN 2003 Conference on Programming Language Design and Implementation (PLDI'03)* (Vol. 38, No. 5, pp. 1–11). ACM Press.

Han, R., Bhatti, S., Carlson, J., Dai, H., Deng, J., Rose, J., Sheth, A., Shucker, B., Gruenwald, C., & Torgerson, A. (2005, August). MANTIS OS: An embedded multithreaded operating system for wireless micro sensor platforms. *ACM/Kluwer Mobile Networks & Applications (MONET), Special Issue on Wireless Sensor Networks, 10*(4), 563–579.

Hill, J. L. (2003). *System architecture for wireless sensor networks*. Unpublished doctoral dissertation, Computer Science Department, University of California, Berkeley.

Hui, J., & Culler, D. (2004). The dynamic behavior of a data dissemination protocol for network programming at scale. In *Proceedings of the 2nd international conference on Embedded networked sensor systems* (Vol. 1, pp. 81–94). ACM Press.

ICL - Imperial College London. (2006). *Ubiquitous monitoring environment for wearable and implantable sensors*. Imperial College London University Web site. Retrieved February 6, 2008, from http://www.doc.ic.ac.uk/vip/ubimon/partners/index.html

Java Tutorials. (2006). *Tutorials trail: The reflection AP*. Retrieved February 6, 2008, from, http://java.sun.com/docs/books/tutorial/reflect/

Jovanov, E. (2006). *WHMS - wearable health monitoring system*. The University of Alabama in Huntsville, Electrical and Computer Engineering department Web site. Retrieved February 6, 2008, from http://www.ece.uah.edu/~jovanov/whrms/

Kara, S., Kemaloglu, S., & Kirbas, S. (2006). Low-cost compact ECG wih graphic LCD and phonocardiogram system design. *Springer Jornal of Medical Systems, 30*, 205–209.

Linz, T., Kallmayer, C., Aschenbrenner, R., & Reichl, H. (2006). Fully integrated EKG shirt based on embroidered electrical interconnections with conductive yarn and miniaturized flexible electronics. In *Proceedings of International Workshop on Wearable and Implanted Body Sensor Networks (BSN'06)* (Vol. 1, pp. 23–26). IEEE Computer Society Press.

Valdastri, P., Menciassi, A., Arena, A., Caccamo, C., & Dario, P. (2004, September). An implantable telemetry platform system for in vivo monitoring of physiological parameters. *IEEE Transactions on Information Technology in Biomedicine, 8*(3), 271–277.

Welsh, M. (2006). *CodeBlue: Wireless sensor networks for medical care.* Harvard University, Division of Engineering and Applied Sciences Web site. Retrieved February 6, 2008, from http://www.eecs.harvard.edu/~mdw/proj/codeblue/

Yang, G. (2006). *Body sensor networks*. London: Springer-Verlag.

KEY TERMS

Application-Oriented Operating System: An operating system that is designed or customized for fulfilling the interests of an application or of a set of similar applications that share a significant part of the requirements. This kind of system must include specific mechanisms and policies, in order to effectively serve the interest of the applications.

Automata or State Machines: A mathematical model that can be used for representing the behavior of a computational system. A state diagram is used for its graphical representation. Basically, an automatum is composed of states that can be graphically represented by circles and edges, which indicate the actions that are responsible by changes of state. In WSNs, usually the states are associated with the tasks, and the edges are directly related to the events that are responsible by the changes in states.

Context Awareness: The capability of computational systems of determining circumstances and scenarios in which its users may be inserted, or simply determining information that is of interest to the system. To do so, the system uses rules (intelligent algorithms) and data that can be supplied by the user itself, or obtained by sensors. These data can also lead to probabilistic information. In this case, the context awareness is not exact, but instead, a hypothesis with an associated probability value.

Multiprogramming: The approach that allows several processes to run simultaneously in a system. In a multiprogrammed sensor node, several programs are simultaneously maintained and managed in the memory by resident software. These programs are organized as tasks. The execution of these tasks follows a policy that is defined during the design of the system, and it is coordinated by a mechanism called task scheduler. In WSNs, there are two kinds of multiprogramming: (i) event-driven run-to-completion single thread approach, and (ii) preemptively time-sliced multithreading model.

Ontology: A data model that represents the entities that are defined and evaluated by its own attributes, and organized according to a hierarchy and a semantic. Ontologies are used for representing knowledge on the whole of a specific domain or on of it. In WSNs, ontologies are included in software architectures as a way of facilitating the search for relevant information. For example, in a BSN an ontology could be used for sensing a specific context (context awareess).

Pervasive Monitoring: The capability of a system to keep itself operating under any condition. These conditions can be related not only to the location (space), but also to the time. They can also be related to technical conditions, such as connectivity, safety, and the amount of energy stored in the batteries.

Preemptive Multitasking: An approach of multiprogramming. The operating system can interrupt the execution of a task and initiate the execution of another, in order to satisfy the restrictions that are defined by the policy that is being used for the scheduling of tasks. It also allows the execution of a task to be suspended or interrupted by a command received from the application.

Quality of Service (QoS): The capability of the system of adjusting itself or of offering mechanisms that allow its adjustment in order to fulfill the requisites that are defined for each application. In WSNs the parameters that are commonly related to the quality of the service that is offered by this kind of system are application lifetime (which is related to energy consumption), connectivity, confidentiality, reliability, bandwidth, and transmission power.

Reflection: A concept related to software: programming languages, operating systems, middleware, and Graphical User Interface (GUI). A reflexive system offers mechanisms that allow its data structures to be inspected and/or modified during execution (at run-time). For this, the system must keep its selfrepresentation that is commonly organized as metadata. In nonreflexive systems, the metadata usually are lost or discarded during compilation, typically when the low-level codes (assembly language) are generated.

Web Services: Modular, independent, selfdescriptive programs that are designed to guarantee interoperability among systems that are developed with different technologies and that interact in a computer network. Typically, Web services are described by using the WSDL (Web Services Description Language), and they use SOAP (Simple Object Access Protocol) for message exchange.

A Prospective Observational Study of the Determinants of Current Cataract Surgical Selection

B. Dhillon
Heriot Watt University, UK

P. Aspinall
Heriot Watt University, UK

T. Aslam
Heriot Watt University, UK

C. Aspinall
Heriot Watt University, UK

P. Halpin
Heriot Watt University, UK

A. Vani
Heriot Watt University, UK

P. Byrne
Heriot Watt University, UK

A. Hill
Heriot Watt University, UK

INTRODUCTION

The principal aim of this research proposal was to develop quality of life and relevant visual function criteria to provide a rational decisional framework to be used as a basis for individual and policy decisions on cataract referral and surgery. There are two key objectives within this overall aim. The first is to identify reliable and effective quality of life and visual function criteria to determine the threshold for surgery for each individual patient in which false positive and false negative errors are minimized. The second is to provide a decisional framework in which risk factors may be incorporated to assess the likely outcome benefits from surgery.

AIMS

The overall aim of the proposed project was the development of criteria, which would provide a rational decision framework as a basis of practical and policy decisions on cataract referral and surgery. The three key objectives within this framework are:

1. The development of reliable and effective criteria to determine the threshold for surgery for each individual patient in which error rates outlined are minimized.
2. To provide a decisional framework in which risk factors may be incorporated to assess the likely outcome benefits from surgery.
3. To produce a measure(s) in which likely outcome benefits (particularly quality of life) from surgery may be prioritized

BACKGROUND

Cataract surgery represents the single greatest highest number of surgical procedures in ophthalmology (Desai, 1999) and its positive health impact was clearly recognized in the UK Government's "Action on Cataract" initiative introduced in 1999 that was designed to streamline the steps, process, and systems of all stages in the patient care pathway. Technical developments in surgery now mean that earlier intervention is much more effective than it was a few years ago and this has lead to increased public expectations to minimize the impact of cataracts on vision related quality of life. Furthermore, the progressive ageing of the population means that there will be an increasing demand for cataract surgery for at least the next 20 years and, unless there are radical innovations to surgical procedure or intra-ocular lens design, it is also likely that referral for post-operative laser capsulotomies will increase proportionally. The clinical context to this project was to investigate factors that determine the process of care early in the cataract patient journey and how this influences surgical selection and subsequent management. Ensuring that there is equitable access to cataract services UK-wide requires some degree of standardization in surgical selection whilst preserving patient-centered care to meet the ophthalmic needs of individuals. Whilst hospital-based care may be more concerned with consequences further downstream in the cataract journey, that is, with increased numbers of patients listed for cataract outstripping surgical capacity, clearly the selection of patients presurgery is critical in the care planning process. The drive to reduce waiting times in cataract surgery has increased over the past several years following the Action on Cataract plan. This has led to initiatives for increasing cataract capacity becoming manifest in a number of public/private initiatives, for example, Netcare and Independent Treatment Centres. Concerns have been raised over "outsourcing" cataract surgical services, effectively unlinking the surgery from pre-assessment selection.

Following the "Action on Cataract" initiative, several fast-tract programmes of direct referral for cataract surgery from optometrists working in the primary care sector have been established (e.g., at the Gloucestershire Eye Unit, Cheltenham General Hospital, and Princess Alexandra Eye Pavilion, Edinburgh) where listing for day-case surgery is largely based on an estimate of potential benefits and complication risks made by the referring optometrist. While this procedure has been shown to work well at a local level where a good working relationship exists between the consultant ophthalmologists and local community optometrists, it has been difficult to implement such programmes on a national scale. This is due to a number of factors, the principal two of which are:

1. Large individual differences between ophthalmologists in what constitute criteria for cataract surgery (Mordue, Parkin, Baxter, Fewcett, & Stewart, 1994), and
2. Large individual differences in the referral criteria for cataract surgery used by optometrists (Latham & Misson, 1997).

Clearly, improvement in referral criteria can only be achieved once there are commonly accepted criteria for surgery. A consequence of these practical problems is that, in some areas of the country, there are high false positive referrals of patients with cataract and nationally there is no estimate of the false negative error rate. These issues are compounded by current practice. At present, the commonly accepted procedure for determining the false positive and false negative error rates is to assess the correctness of the referral in terms of whether an ophthalmologist will diagnose a patient as having a level of cataract that is visually disabling. Given the individual differences within ophthalmology in applying criteria for cataract surgery, and given that the appropriate validation criterion is not the diagnosis but whether the patient is likely to benefit from surgery, there is an urgent need to develop commonly agreed upon criteria used both by ophthalmologists and optometrists that will comprise reliable and effective measures and procedures for determining the threshold for cataract surgery for each individual patient.

The object of cataract surgery is to improve visual function and many reports demonstrate that the vast majority of patients benefit from surgery (Minassian, Rosen, Dart, Reidy, Desai, & Sidhu, 2001). For many years the emphasis has been on Snellen visual acuity performance both for referral and as an outcome indicator (Ruit et al., 2007). However, several studies have shown that the disabling effect of cataract is largely due to a reduced contrast in the retinal image due to intra-ocular light scatter. Not surprisingly, these studies on patients with cataract have shown that visual acuity measures correlate poorly with visual symptoms (Rubin, Adamsons, & Stark, 1995; Steinberg et al., 1994) and that changes in contrast sensitivity, especially in the presence of glare, give better predictions of the difficulties that patients face in daily situations (Elam, Graney, Applegate, Miller, Freeman, & Wood, 1988). Moreover, where cataract co-exists with another vision reducing pathology such as age related macular degeneration (AMD) (Armbrecht, Findlay, Kaushal, Aspinall, Hill, & Dhillon, 2000) or glaucoma (Nelson, Aspinall, Papasouliotis, Worton, & O'Brien, 2003), we have shown that functional measures of contrast sensitivity in the presence and absence of glare are more effective predictors of quality of life indicators than visual acuity.

A number of approaches have been developed that aim to assess visual symptoms and functional disability. These measures are able to identify those patients with significant disability due to their cataract and may assist in deciding whether cataract surgery is indicated. However they are poor at discriminating those patients with mild to moderate visual disabilities and in addition they have not been evaluated postoperatively. It has also been suggested that measures of vision related aspects of quality of life such as the VF-14 (Steinberg et al, 1994), the Activities of Daily Living Scale (Mangione, Phillips et al., 1992), or the Catquest Questionnaire (Lundstrom, Stenevi, & Thorburn, 2001) might be better indicators of the benefit of cataract surgery. It is worth noting however that the main applications of questionnaires are grounded in sampling theory so that the results between groups of patients can be compared and they are therefore more unreliable and less useful as a means of assessing individual needs (Armbrecht, Dhillon, & Aspinall, 2001). It remains the case that a properly validated scale of vision related quality of life, or a combination of such a scale and a relevant visual function test that may be used to assess who would benefit from cataract surgery, has still to be developed. The background literature suggests that the visual_function test will be one that demonstrates the disabling effects of veiling glare on contrast threshold performance. Unfortunately, the lack of any standardization and effective calibration of glare tests such as the Brightness Acuity Tester has meant that this approach has not gained clinical acceptance. At this point we had two choices. The first would be to use a functional test that is based on a practical daily task such as reading. We had developed such a test as one of the outcomes of our recent research on quality of life in patients with cataract having the comorbidity of AMD or glaucoma. This test was developed as part of an EPSRC award on practical problems arising from glare and could provide an estimate of the disabling effects of cataract (Hill & Aspinall, 2002). It has the advantage of being a simple and quick to use contrast visibility indicator. However, at the start of the project, the contrast visibility indicator test had not been evaluated. The second option arose from a research link with Dr. Tom Van den Berg of the Institute of Ophthalmology in Amsterdam. He had developed a new instrument (the C-quant) for assessing light scatter in the eye and presented early evaluations of the equipment at the ARVO conference in Florida in April 2004. Following discussion, it was decided that it was of mutual benefit to use this proven measure of light scatter within the project.

The problems relating to the lack of agreement on referral or intervention criteria for capsulotomies from postoperative capsule opacification are similar to those for the primary referral for cataract surgery itself. The development of any referral system incorporating decisional support on visual disability, quality of life, and patient awareness of risk for cataract surgery will, therefore, have practical relevance to establishing criteria for the management of secondary cataract from postoperative capsule opacification.

It was, of course, recognized that not all decisions about whether to refer for cataract surgery can be based solely on symptoms and a measure of reduced visual function. There are many risk factors that need to be considered by the surgeon and, consequently, these risk factors also need to be recognized by the referring optometrist if false positive referrals and unrealistic patient expectations are to be avoided. In order to aid estimates of surgical outcome, Moorfields Eye Hospital has recently taken the initiative to examine ways in which the risks of complications from cataract surgery may be used to prioritize patient management (Muhtaseb, 2004). They have highlighted the risks of ocular comorbidity but there are also several important nonocular risks. For example, the elderly and those with persisting cognitive impairments often do not show any recovery in daily activities post operatively from cataract surgery (Elam et al., 1988). It was desirable, therefore, that effective decisional support strategies be developed that will assist in the cataract referral process and that these will have wide professional acceptance. There are now many examples of the way in which decision models and strategies can be used to assist in clinical decision making in ophthalmology (Massof, 2002; Piermarochi et al., 2006) and we explored ways in which these could gain practical acceptance for the more effective management of cataract. A recent quality of life scale (VSQ, O'Donovan et al.) was included to complement the traditional VF14 (Sparrow, Bron, Brown, Ayliffe, & Hill, 1986).

Whilst Mordue et al. (1994) have used consultant interviews to explore variation in cataract surgical selection (Mordue et al., 1994), our approach using both interview and assessing selection in the *same* group of professionals is unique. In addition, we wished to extend further recent evidence from Lash, Prendiville, Samson, Lewis, Munneke, and Parkin (2006, pp. 464-467) that the interface between optometrist and patient is key to refining surgical selection.

PATIENTS AND METHOD

In developing the decisional framework, the relevant criteria and their associated Quality of Life (QoL) scales were piloted within the ophthalmological, patient, and optometry communities in Edinburgh. The main source of data collection was from the six cataract referral clinics at the Princess Alexandra Eye Pavilion in Edinburgh by the research team experienced in visual psychophysics, cataract grading, straylight measurement, and QoL measurement masked to the findings of referring optometrist and supervising ophthalmologist. Ethical approval was granted for the study and all patients had fully informed consent. The study dataset included:

a. **Visual function**—Snellen acuity, Logmar distance, and near acuity, Pelli Robson contrast sensitivity.
b. **Clinical**—objective grading of cataract morphology as assessed by the Oxford clinical cataract and grading system (Steinberg et al., 1994), Stray light meter test (C-quant).
c. **Quality of life**—VF14, Massof QoL scales, and VSQ for visually impaired people (Sparrow et al., 1986) developed at the Wilmer Eye Institute at Johns Hopkins University.

Within the constraints of the study period, follow up tests were performed at 3 months after surgery. The main interest here was a comparison of pre and post op quality of life indicators.

Following all data gathering, the research aim was to draw up a decisional framework and set criteria to summarize the views of the groups as a basis for further discussion. This 1-year study aimed to produce a prototype decisional framework. The longer view of the research will attempt to derive a consensus view initially using the Delphi technique, before summarizing opinion. The opinion of an expert ophthalmology group in conjunction with those of patients would provide a consensus towards the formation of an established norm (i.e., a "gold standard").

The data obtained from these sources will be used in conjunction with QoL scale questionnaire development for decision analysis. For example, the relationship between visual function data and the degree of change in scale items pre and post surgery will be examined. We need to know what level of functioning is impaired prior to surgery and what say 5% of improvement really means in relation to quality of life. Finally a prototype numerically based profile scale for quality of life outcomes would be drawn up for test evaluation.

The value of the research to Public health and patient care could be summarized as:

- Improving relevant referral criteria that relate to decisions about the likely benefit of cataract surgery for an individual.
- Use of common criteria nationally.
- Reduction of false positive error rates.
- Quantification of false negative error rates.
- Consensus of decisional criteria within and between the relevant professional groups in primary and secondary health care.
- Improved cost effectiveness of ophthalmology resources.
- Improved effectiveness of managing patient expectations from cataract surgery.
- Improved Quality of Life for patients.

RESULTS

Main Data Collection

105 patients referred for cataract surgery to the Princess Alexandra Eye Pavilion were given a series of tests. The clinical tests including the stray light meter test;

a series of tests of visual function including the LogMar visual acuity test and the Pelli Robson contrast sensitivity test; and interviews using three quality of life scales as described above. A brief overview of the demographic characteristics of the patients is shown in Figures 1 and 2.

Demographics

The mean age is 72 years with a standard deviation of 11 years. In addition there was an even gender split with 55% of patients being female and 45% male, while 78% of patients were referred for surgery. Of these latter patients, 77% of the patients listed for surgery were first eye patients. Around 20% of referrals were not listed for surgery.

Visual Function of the Total Patient Group

In order to help interpretation of this data for Snellen acuity, which is the most widely known, the two highest frequencies within which the mean visual acuity lies, correspond to acuities of 6/9 and 6/12 with 6/12 .

This distribution is at the better end (and therefore the more controversial and problematic end) of the acuity range where we had wanted it to be. Decisions about cataract surgery at poorer acuities (other things being equal) are obviously not problematic and therefore not the main focus of this study.

Listed vs. Non Listed Patients: The Differences

As the initial purpose of this project is to understand the factors which are currently used in practice to discriminate those who are listed for surgery from those who are not listed, the initial analysis was to determine these discriminating variables. In Table 1, a nonparametric statistical test (Mann Whitney U test) shows the discriminating variables for right eyes between those listed and not listed at a probability <0.05. The variables are presented in 3 major categories—first clinical data; then visual functional data; and finally quality of life data.

Out of the full variable list, the following variables were significantly different (1-tailed test) between patients who were listed or not listed.

Table 1. Difference between listed and non listed patients

Category	Variable	Mann Whitney	Probability
Clinical	Cataract asc	2.0	0.04
	Cataract waterclefts	1.8	0.06
	Cataract brunescence	2.6	0.009
	Cataract white scatter	2.2	0.02
	Cataract psc	2.1	0.03
	Cataract fibrefolds	1.7	0.08
	Iris colour	2.0	0.03
	Stray light meter	3.2	0.002
Visual Function	Snellen	4.1	0.0001
	LogMar (distance)	4.1	0.0001
	Log(pin)	2.6	0.009
	Pelli Robson	3.6	0.001
	PelliRobson (near)	3.6	0.001
Quality of Life	Massof (Gardening)	2.3	0.02
	VF14 Driving (diff)	1.8	0.08
	Vsq Glare (diff)	2.2	0.02

Right Eye

In the clinical data, ocular comorbidities are not significantly different between those listed and not listed which may reflect the small number of comorbidities in the group. From the range of cataract features ASC is significant while water clefts is just beyond the boundary of significance. PSC is significant while brunescence is highly significant and white scatter is also significant. These latter two are the most significant discriminating variables from the set of cataract graded features. Iris colour is also just significant. In the visual function category, Snellen acuity, LogMar distance acuity (including pinhole method) is highly significant, as is Pelli Robson contrast sensitivity. LogMar near reading acuity and the stray light meter are also highly significant discriminators between those listed and those not listed for surgery.

When we examine the QoL scales "gardening" reaches significance as does driving and glare. It should be noted that quality of life scales would be more likely to relate to binocular visual function rather than criteria for single eye surgery.

A particular focus of this research was to look at criteria for surgery in patients with relatively good visual acuity. The Snellen acuity of 6/12 is seen as a benchmark by ophthalmologists for different reasons including its critical role in determining whether someone is fit to drive. Surgery on patients with visual acuity better than 6/12 is currently a topic of much debate. In Table 2, the criteria discriminating those with acuities 6/12 or better who were listed or not listed for surgery is given.

Significant differences for patients listed or not listed.

Right Eye

Again within the patient group with better acuities, a similar pattern of discriminating variables between those listed and not listed for surgery is seen.

As an example for the subgroup of patients with acuities better than 6/12, the difference between those listed and not listed for surgery is shown below in Figure 1 for the variables brunescence, white scatter, LogMar acuity, and Pelli Robson.

Figure 1 shows two of the clinical and visual functional variables clearly discriminating between patients who are listed or not listed with acuities better than 6/12.

It is also interesting that for both right and left eye data there is no difference between the significant variables for listing for the six consultant teams.

Quality of Life Scales

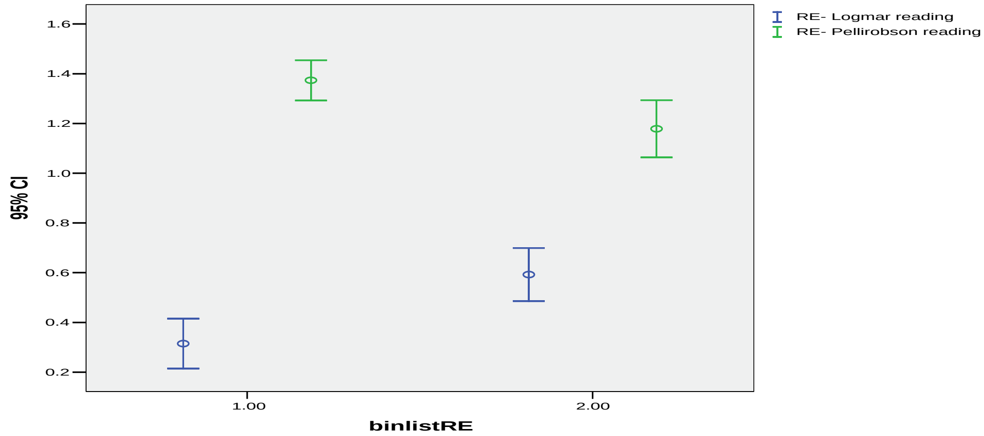
Quality of life, which is increasingly seen as a basis for all clinical intervention, has been measured using

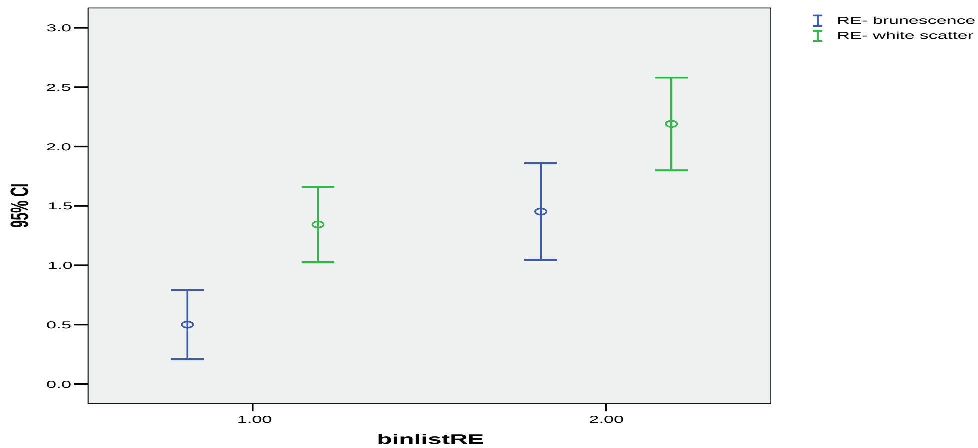
Table 2. Subgroup of patients with acuities better or equal to Snellen 6/12

Category	Variable	Mann Whitney	Probability
Clinical	Cataract cortical spokes	2.1	0.04
	Cataract Brunescence	3.2	0.001
	Cataract white scatter	3.1	0.002
	Stray light meter	3.3	0.001
Visual function	LogMar distance	3.4	0.0001
	LogMar (pin)	2.0	0.03
	Pelli Robson	2.6	0.01
	LogMar (near)	2.6	0.007
Quality of life	Vf14.10 sport	2.0	0.04
	Vsq 11 (recognising faces)	1.9	0.05
	Vsq12 (glare)	2.4	0.02
	Vf14.8 (writing)	1.8	0.08

P

Figure 1. Differences in LogMar acuity, Pelli-Robson contrast sensitivity, brunescence, and white scatter for listed (2) and not listed (1) patients





three different scales in this study. The VF-14 scale has been widely used while the VSQ scale and the Massof scale are more recent. The data presented above has indicated that there are elements of the scales which are different between those listed and not listed.

VF. 14 Scale

A frequency plot of responses on the vf14 scale showed half the sample not responding to questions 3,7, 9, and 10 (large print, fine handwork, card/board games, and sports, respectively). A factor structure of the remaining variables (excluding driving) showed from a principal component analysis a two factor solution which accounted for 57% of the variance in the data.

VSQ Scale

A similar analysis was carried out on the Vsq scale. The third question of the scale “which is your bad eye” was not included in the following analysis. Here, five factors accounted for 70 % of variance in the data.

The correlations between acuity data and quality of life scales are significant but moderate. In other words there is no question that binocular visual acuity on the Logmar or Pelli Robson tests is highly significantly related to quality of life scores on VF14 and Vsq scales. However the highest relationships in the table are correlations in the range 0.3 to 0.4 which means that the percentage of common variance between acuity and quality of life measures is between 9% and 16%, leaving at best around 84% of the variance unaccounted.

This is a case therefore of statistical significance with limited applied or practical benefit.

A clearer picture of the relationship between binocular visual function and quality of life is provided by mapping the two quality of life scales VF14 and Vsq against the two binocular visual function measures of Logmar and Pelli Robson contrast.

Two interesting aspects emerge. First, the relationship of the VF14 scale with the two measures of visual function is linear in both cases. On the other hand, the relationship of the Vsq scale is nonlinear and for the case of the Pelli Robson contrast test and the LogMar test, shows a threshold score beyond which quality of life falls. This requires further exploration but if confirmed provides useful evidence based practical criterion before which quality of life is relatively stable and unaffected and beyond which quality of life deteriorates. The Logmar criterion is around a score of 3 (or 0.19-0.30 on the test chart). The PelliRobson criterion is at a score of 2 (or 1.36-1.5 on the test chart). The nonlinear aspect of Vsq accounts for its slightly lower correlation than VF14 with visual function.

Quality of Life: Pre Operative vs. Post Operative

Perhaps the most important outcome measure in the study is the difference in quality of life indicators between a patient's situation before and after surgery. The number of patients available for this comparison was much smaller than the sample undergoing surgery. This was due to the limitation of the study (one year duration) and the waiting times before surgery and 3 months after surgery for follow up quality of life data. However, in spite of the reduced numbers available, there were highly significant differences in t scores in 10 of the 12 quality of life scales between the situation before and after surgery, with the differences being positively linked to surgical intervention. On quality of life indicators, surgery was clearly perceived as beneficial. A comparison of individual patient data shows that for all patients the score if not in the positive direction remained the same post surgery. In no instance did a patient report a worse quality of life score post surgery.

Clearly there is a high correlation between acuity level and difference in acuity between eyes – r=0.758 and r=0.817 for right and left eyes, respectively.

DISCUSSION

Towards a Prototype Decision Model

One of the central aims of the study was to make recommendations towards a prototype decision model. We are now in a position to do this but we need to emphasize two issues. First, there has been a significant limitation on the numbers of patients available for study in spite of accessing all cataract clinics in the Princess Alexandra Eye Pavilion. This limitation has been particularly severe on the follow up patients—for example, given a waiting time prior to surgery and a post operative minimum of 3 months post surgery we have only been able to see 28 patients before and after surgery. Nonetheless the significance of the findings point to particular conclusions. Second, the study has been based at one particular hospital across five consultant teams. We do not know how representative these consultants are with respect to the UK in general although qualitative interviews (results reported elsewhere) and the questionnaire data suggest that they may be typical of UK consultants. We need therefore to test the prototype more widely across the UK.

There were two initial tasks:

a. To find variables which best discriminated in current practice between those patients referred and not referred for cataract surgery.
b. By using the outcome benefits to patients from cataract surgery, to identify a rational basis for changing current practice.

We have identified the variables under (a) and will use these in the decision model. However, the issue under (b) of changing current practice raises an important question. The Edinburgh data clearly shows that current practice is working in delivering perceived benefits to patients. While a small number of patients indicate no improvement in daily tasks post surgery, there is no evidence of patients regretting their decision to go ahead with surgery or of reporting more negative quality of life scores post surgery. We have no information on the satisfaction or otherwise of patients who were refused surgery. In other words, given we have a group of very satisfied patients, we have no grounds from users of the service for changing current local practice.

We have approached the research by assuming that in ophthalmology in general, there are three categories of variables relevant to any appropriate decision model. These categories are clinical state, visual function, and quality of life. We believe it is important to preserve these categories in the final prototype model for cataract surgery. In the analysis that follows, eyes are considered as units for analysis rather than persons. So a particular eye has a functional and clinical profile which provides the evidence for or against surgery. The quality of life evidence is more problematic as its main basis is binocular visual function. In fact, in the study there are significant (though moderate) correlations between binocular vision and quality of life. The quality of life link with monocular vision would be expected to be weaker as the main criterion for quality of life would be vision in the better eye. As a consequence, quality of life criteria would be less likely to discriminate between referrals in current practice (and therefore be shown in regression analysis) irrespective of their low frequency of use.

Another important issue is the acuity level for surgery. We have taken the view that surgery on Snellen acuities greater than 6/12 is noncontroversial and consequently of little research interest. Far more important is the discovery of a rational basis for surgery at the better end of the acuity range.

A final criterion in the decisional model was ease of use in practice so that ophthalmologists and optometrists could be provided with relatively straight forward criteria for referrals.

ANALYSIS FOR PREDICTING CURRENT PRACTICE

The results in Tables 1 and 2 show the variables which are significant discriminators of current referral practice under the three categories described. The subset of these, which were most highly significant, was used in regression analysis. These were:

i. Clinical State
    Brunescence
    White scatter
    Stray light meter reading
ii. Visual Function
    Snellen acuity
    LogMar acuity
    Pelli Robson contrast sensitivity
iii. Quality of life
    Summed scores across vf14 questionnaire
    Summed scores across the vsq questionnaire

In addition, difference scores between right and left eyes were calculated for the three tests of visual function.

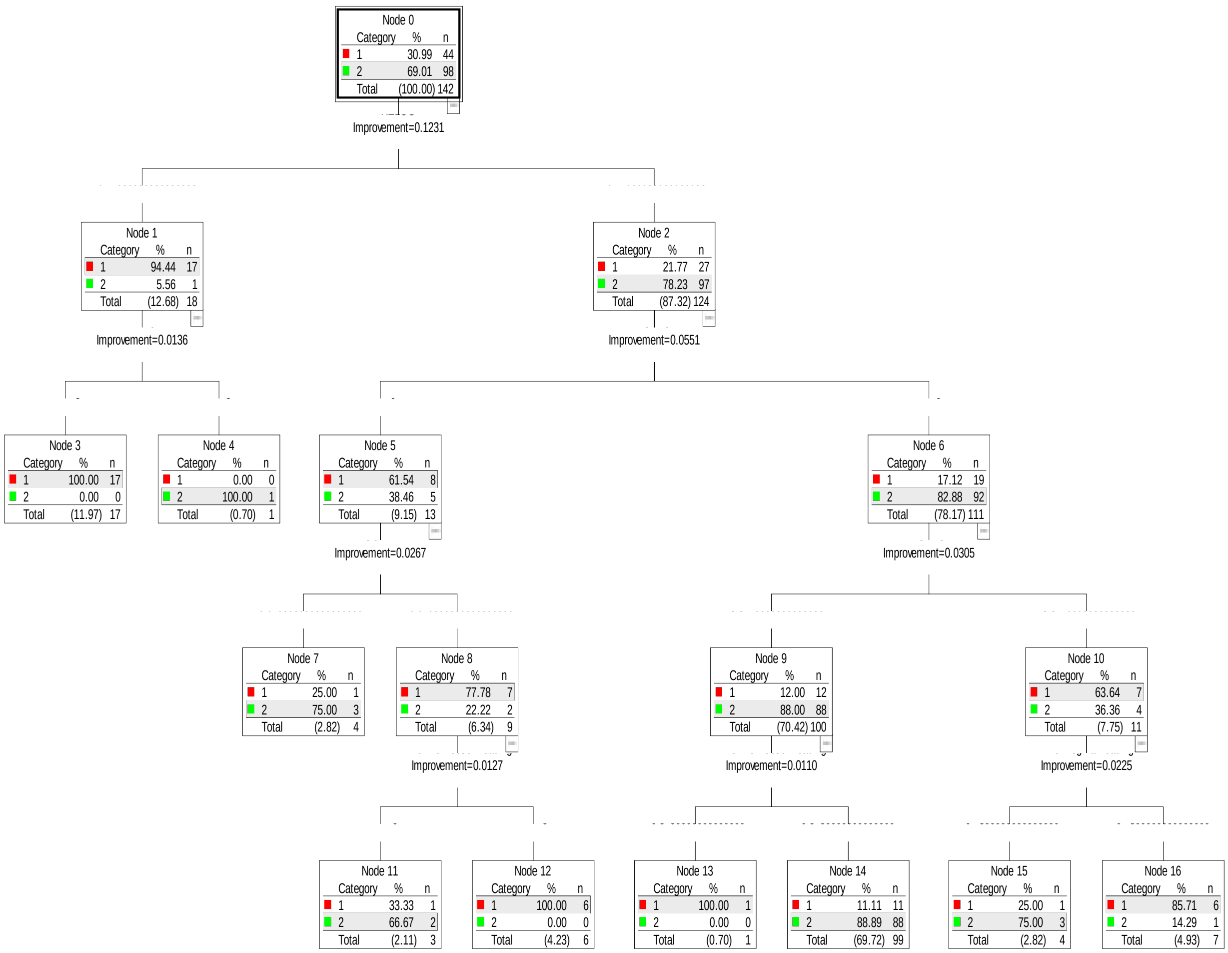
The chosen analysis was the decisional package Answer Tree—a module of SPSS. This has several advantages. First, it is an exploratory decisional which gives flexibility to the user. Users can manipulate the model to provide alternative criteria at different nodes of the tree and receive feedback on the impact of this change to the tree as a whole. This feature of local control should appeal to consultants and to local authorities with special concerns or needs. The analysis also has the advantage of being a nonparametric form of analysis and so able to cope with a variety of non normal distributions and nominal data. In addition it provides an optimal basis for a sequence of decisional nodes, and identifies the most efficient criterion point for critical paths at any node.

Predicting current practice with target variable surgery listing (all patients) and the algorithm classification and regression is shown in Figure 2. The strongest predictor is that at the first node of the tree and is LogMar acuity. By following the change in frequencies in each box as you move down the tree, it is possible to see the impact of the variable at that node on classification. Although not shown, the software provides a rank order of alternative predictors at each node of the tree and a measure of relative improvement in classification. Finally a misclassification matrix gives the overall accuracy of the model.

In summary, a tree made up of binary splits gives a good prediction of current practice. Key discriminators near the top of the tree are LogMar acuity, brunescence, and the stray light C-quant measure. Difference between eyes and binocular vision appear near the foot of the tree.

When a regression analysis is run on those patients with better acuities (i.e., <=6/12 Snellen), the relative importance of both the stray light measure and Pelli Robson contrast sensitivity as independent predictors of patient listing for surgery increases. In fact a simple three level tree with stray light at the first node and difference in contrast sensitivity between eyes at the second node produces a classification matrix at around 90% accuracy.

Figure 2. Regression tree for all patients



Analysis for Predicting Quality of Life

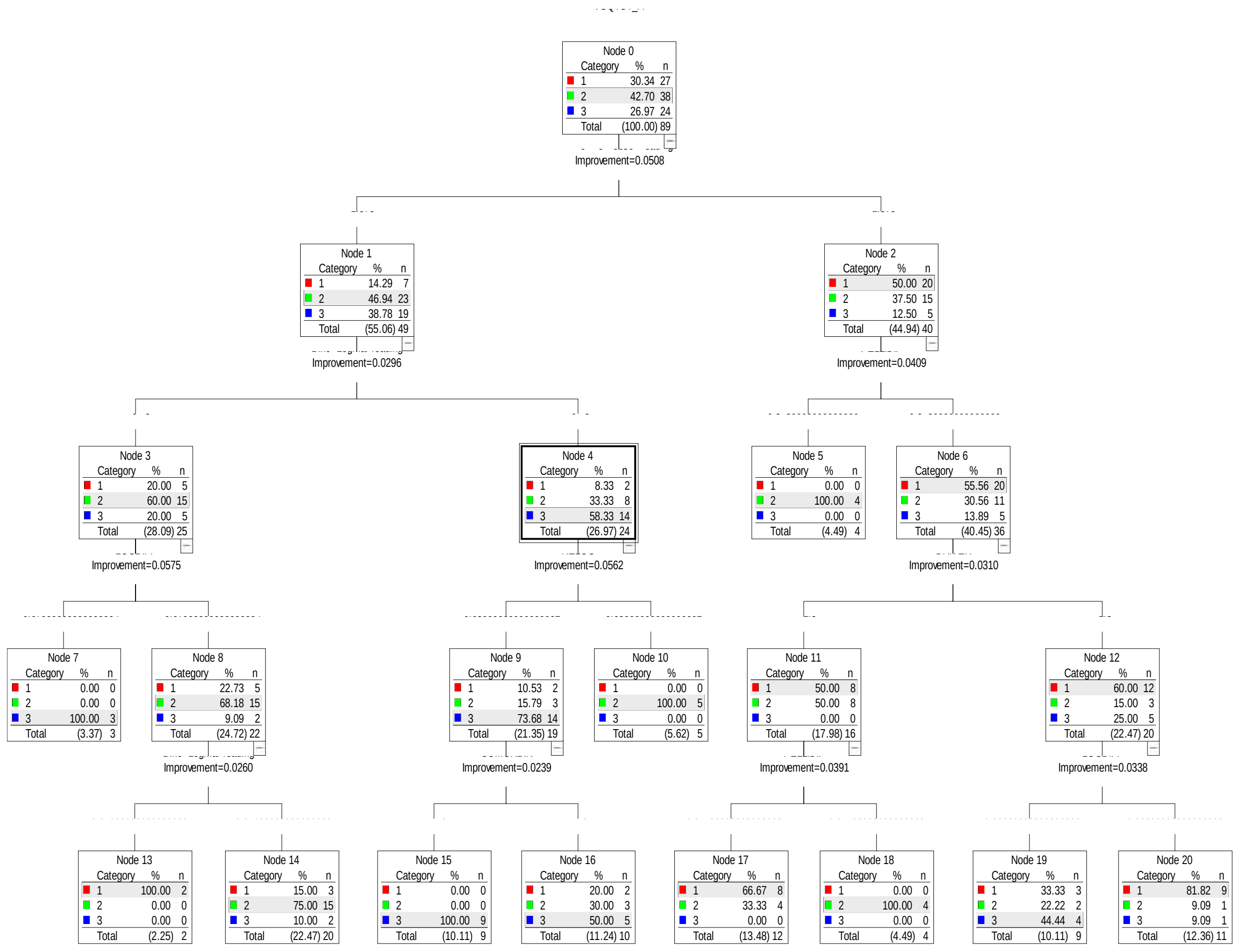
When the key variable for referral is changed to Quality of Life rather than current practice, Quality of Life is best predicted by binocular vision. In the case of the Vf14, scale binocular LogMar appears at the top and lower down the tree, and in the case of the Vsq scale, it is the binocular Pelli Robson contrast sensitivity test which is the best predictor. Once again brunescence and stray light measures feature very strongly.

Similarly if we examine the predictors of quality of life for the top acuity group (<=6/12 Snellen), then as in the total group of patients it is binocular contrast sensitivity which is the main discriminator for quality of life. This is shown in the regression tree in Figure 3. It is noticeable that difference between eyes is featuring strongly as an important variable in the trees. Driving and comorbidity also feature as relevant.

In summary, the regression trees show high prediction for current practice for all patients referred and for those in the top acuity group. When quality of life is made the key dependent variable, the model prediction is less good and binocular contrast sensitivity appears more strongly as a relevant predictor.

At this stage, we need more evidence from a wider based study of hospital practice. It seems to us unlikely that the relevant parameters in cataract referral will differ between hospitals, but it seems highly likely that the weightings given to these parameters may well differ between centers. As a consequence, until we receive further feedback on the Edinburgh weightings, the regression weights are not built into the simplified decisional outline scheme given below.

Figure 3. Regression tree for predicting quality of life for the top acuity patients



FUTURE WORK

Simplified Decision Structure for Cataract Surgery

Notes

1. RED indicates Refer (or for ophthalmologists) list for surgery, AMBER indicates maybe, and GREEN indicates no basis for referral on that measure.
2. The 3-way split (1, 2, 3) on scales would be based on proportions of patients in the best third (1), middle third (2), and worst third (3) of each scale.
3. Note that the risk scale is inverted and that the lower risk categories (1 and 2) are both given amber. In other words, low risk cannot of itself justify referral.
4. Similarly for patients views, which no matter how keen for surgery are not sufficient in themselves.
5. Any one RED appearing in the profile of a patient indicates refer. The challenge is to find the threshold for how many AMBERS indicates referral. We assume this will be arrived at in consultation with ophthalmologists (we suggest 3 AMBERS=refer).
6. The model assumes that all three major categories, and all measures are equally important in arriving at a decision. This is not true with respect to analysis of current practice but may be a useful starting position. The same assumption underlies tertiary splits for criterion points on each scale.
7. We have not as yet gone back to the data and looked at the consequences of implementing this or discovering the number of AMBERS meeting current practice.

Box 1. Simplified decision structure for cataract surgery

Vision		**Decision**
Snellen	worse than>6/12	RED
LogMar	1	GREEN
	2	AMBER
	3	RED
Pelli Robson	1	GREEN
	2	AMBER
	3	RED
Clinical		
Brunescence/White Scatter/Stray Light		
	1	GREEN
	2	AMBER
	3	RED
Risk	1	AMBER
	2	AMBER
	3	GREEN
QoL		
VSQ Questionnaire	1	GREEN
	2	AMBER
	3	RED
Driving	No	GREEN
	Yes but not essential	AMBER
	Yes and essential	RED
Patient wish (or anxiety)	Undecided (not high)	GREEN
	Keen	AMBER

CONCLUSION

This study has described the determinants of current cataract surgical selection in a large teaching hospital. However, there is scope for improved concordance in referral criteria and selection. There is a need to expand the dataset to cover a wider spread of consultants working in different geographic locations in both teaching units and district hospital which tend to serve urban and rural populations respectively.

In addition, we have proposed a simplified yet flexible decision structure for cataract surgery designed to improve both equity of access to cataract surgery and increase the efficiency of the "cataract journey" by reducing false positive referrals from optometrist to hospital specialist. There is an opportunity to maximize further funding support by addressing this issue Scotland-wide in parallel with the Centre for Change and Innovation Programme for improving cataract services. With additional funding, we would aim to carry out a multisite study in Scotland to:

1. Test the decision model proposed,
2. Expand the dataset evaluation in second eye surgery,
3. Gather follow up data over a 2 year period.

ACKNOWLEDGMENT

The authors gratefully acknowledge the generous support of a BUPA Foundation Award, which funded the study, and to the Princess Alexandra Eye Pavilion consultants, nurses, and patients who made this study possible.

REFERENCES

Armbrecht, A., Dhillon, B., & Aspinall, P. (2001). Visual and health related quality of life measurements. *Continuing Medical Education Ophthalmology, 5,* 8-12.

Armbrecht, A., Findlay, C., Kaushal, S., Aspinall, P., Hill, A., & Dhillon, B. (2000). Is cataract surgery justified in patients with age related macular degeneration. *British Journal of Ophthalmology, 84*, 1343-1348.

Desai, P., Minassian, D. C., & Reidy, A. (1999). National cataract surgery survey 1997-1998. *British Journal of Ophthalmology, 83,* 1336-1340.

Elam, J., Graney, M. J., Applegate, W.B., Miller, S. T., Freeman, J. M., & Wood, T. O. (1988). Functional outcome one year following cataract surgery in elderly persons. *Journal of Gerontology, 43*, 122-26

Hill, A., & Aspinall, P. (April 2002). *Lighting design and visual impairment*. Paper presented at the CIE meeting in the Royal Society London.

Lash, S. C., Prendiville, C. P., Samson, A., Lewis, K., Munneke, R., & Parkin, B. T. (2006). Optometrist referrals for cataract and action on cataracts guidelines: Are optometrists following them and are they effective? *Ophthalmic Physiology for Optometrists, 26*, 464-467.

Latham, K., & Misson, G. (1997). Paterns of cataract referral in the West Midlands. *Ophthalmic Physiology for Optometrists, 17,* 300-306.

Lundstrom, M., Stenevi, U., & Thorburn, W. (2001). Quality of life after first and second eye surgery; 5 year data collected by the Swedish National Cataract Register. *Journal of Cataract and Refractive Surgery, 27*, 1553-1559.

Mangione, C. M., & Phillips, R. et al. (1992). Development of the activities of daily vision scale. A measure of visual functional status. *Medcare, 30*, 1111 -1126.

Massof, R. (2002). Measuring low vision rehabilitation outcomes with the NE1 VFQ-25. *Investigative Ophthalmology and Vision Sciences, 43*, 2859-2868.

Minassian, D., Rosen, P., Dart, J., Reidy, A., Desai, P., & Sidhu, M. (2001). Extracapsular cataract extraction compared with small incision surgery by phacoemulsification: A randomized trial. *British Journal 0f Ophthalmology, 85,* 822-829.

Mordue, A., Parkin, D., Baxter, C., Fewcett, G., & Stewart, M. (1994). Thresholds for treatment in cataract surgery. *Journal of Public Health Medicine, 16*, 393-398.

Muhtaseb, M., Kalhoro, A., & Lonides, A. (2004). A system for preoperative stratification of cataract patients according to risk of intraoperative complications: A prospective analysis of 1441 cases. *British journal of Ophthalmology, 88,* 1242-1246.

Nelson, P., Aspinall, P., Papasouliotis, o., Worton, B., & O'Brien C. (2003). Quality of life in glaucoma and its relation to visual function. *Journal of Glaucoma, 12,* 139-150

Piermorochi, S., Sartore, M., Bawdello, F., Lanzetta, P., Brancado, P., Garattini, L. et al. (2006). Quality of vision: A consensus building initiative for a new ophthalmologic concept. *European Journal of Ophthalmology, 16,* 851-860.

Rubin, G. S., Adamsons, I. A., & Stark, W. J. (1993). Comparison of acuity, contrast sensitivity and disability glare before and after cataract surgery. *Archives of Ophthalmology, 111, 56-60.*

Ruit, S., Tabin, G., Chang, D., Bajracharya, L., Kline, D. C., Richheimer, W. M. et al. (2007). A prospective randomized clinical trial of phacoemulsification vs. manual sutureless small inscision Wxtra Capsular cataract surgery in Nepal. *American Journal of Ophthalmology, 143,* 32-38.

Sparrow, J., Bron, A., Brown, N., Ayliffe, W., & Hill, A. (1986). The Oxford clinical cataract classification and grading system. *International Ophthalmology, 9*, 207-225.

Steinberg, E., Tielsch, J., Schein, O. D., Javitt, J. C., Sharkey, P., Cassard, S. D. et al. (1994). The VF-14. An index of functional impairment in patients with cataract. *Archives of Ophthalmology, 112,* 630-638.

KEY TERMS

Brunescence: Reflection of cataract density.

Cataract Surgery: Surgical operation.

False Positives/Negatives: Reflection of accuracy of collected data.

Glare: Measure of light scatter.

Laser Capsulotomies: Creation of opening in capsular bag.

Logmar Acuity: Logarithmic acuity measure.

Moorfields Eye Hospital: UK Ophthalmic hospital.

Pelli Robson Contrast Sensitivity: Indication of how light scatter affects visual function.

Stray Light Measurement: Measure of glare.

Snellen Visual Acuity: Graduated measure of vision.

P

Prospects for Development of Shock Wave Therapy

B.A. Garilevich
Central Clinical Air Force Hospital, Russia

Y.V. Kudryavtsev
Central Clinical Air Force Hospital, Russia

Y.V. Olefir
Central Clinical Air Force Hospital, Russia

Y.V. Andrianov
Central Clinical Air Force Hospital, Russia

A.E. Rotov
Central Clinical Air Force Hospital, Russia

INTRODUCTION

Nowadays the world medical practice widely uses the method of extracorporeal shock wave lithotripsy for treatment of patients with urolithiasis bay means of pulverizing action of focused shock waves (FSW). In the early 1980s in Germany, the Dornier Medical Systems Company fabricated the first lithotriptor based on the electrohydraulic principle of shock wave generation. By 1988, practically all countries had applied lithotriptors developed by Dornier Medical Systems. Later, more than 15 countries started to fabricate their own lithotriptors, which use various methods of FSW generation, such as electrohydraulic, electromagnetic, and piezoelectric. These countries were Germany, USA, Russia, China, Italy, France, Israel, Turkey, Slovakia, Poland, Austria, Sweden, and so forth.

However, the given method of the physical effects can be used not only for the extracorporeal fragmentation of calculi. The biological effects of shock waves presume a perspective of their use for treatment and aftertreatment of the patients with other disorders and pathological conditions using their stimulatory and inhibitory actions.

BACKGROUND

The method of shock wave therapy (SWT) originates in the late 1980s and in the early 1990s. A number of investigations have proved that the influence of shock waves accelerates callus formation and fracture healing, faster healing of skin wounds, as well as intensifies the regenerative and reparative processes and stimulated metabolism in tissues and cells (Garilevich, Kudryavzev, & Dzeranov, 2003; Garilevich, Zakharov, Kudryavzev, & Kirpatovsky, 1997; Haupt, 1997; Valchanov & Michailov; Zubkov, Garilevich, & Olefir, 2005).

In 1988, SWT was for the first time successfully applied for treatment of fractures (Valchanov & Michailov). Thereafter, the method has been widely used in orthopedic and traumatic surgery and sports medicine. With accumulation of experience, the scope of application of this method has been gradually extended. So, its efficiency in treating Peyronie's disease, neuromuscular dysfunction in children with spastic cerebral paralysis (Siebert & Bush, 1997) has been proved. The investigations of the last years have exposed a perspective of cardiologic use of SWT at patients with ischemic heart disease (Nishida et al., 2004). The basic clinical effects of shock waves are: analgesic action, activation of microcirculation and neonangiogenesis, stimulation of the metabolic processes, resorption of healing connective tissue, anti-inflammatory, and antibacterial action (Garilevich, Kudryavzev et al., 2003; Garilevich, Zakharov et al., 1997; Gerdesmeyer et al., 2005; Haupt, 1997; Kudryavzev, Garilevich, Kirpatovsky, & Olefir, 2003; Mironov, Vasiljev, & Burmakova, 1999; Nishida et al., 2004; Zubkov, Garilevich, & Olefir, (2005).

Nowadays, there are a number of issues about the use of the inhibitory action of shock waves. The methods of cancer treatment through the influence upon them by ultrasonic waves of high intensity (U.S. Patent No. 3735755, 1975; U.S. Patent No. 4441486, 1986), and focused shock waves (W.O. Patent No. 86/05104, 1986) are well known. The cytotoxic action of shock waves upon cancer cells (Eisenberger & Miller, 1999; Kohri et al., 1990; Mastikhin, Nicolin, Teslenko et al., 1995) has been also described. The ability of FSW, besides the mechanical damage of cells, to break sulfide bonds and to form highly reactive free radicals makes possible the participation of the free radical mechanism in changing the membrane permeability of cells (Mastikhin et al., 1995).

However, these methods have a restricted application due to a high negative pressure of a shock wave impulse (SWI) provokes cavitation processes in the field of the diffusion of a wave beam that also damages healthy tissues and causes a decrease in the efficiency of tumor growth suppress.

Disadvantages of the known methods have been eliminated in the new offered method of treating malignant neoplasms based on a synchronous influence of SWI and pulsed electrical field of nano or microsecond range of duration (Andriyanov et al., 1999).

Nowadays, various firms produce equipment to be used for the treatment of patients with orthopedic disorders applying FSW. Disadvantages of the given equipment result from the fact that the parameters of SWI do not differ from those applied to the extracorporeal fragmentation of urinary calculi. That is to say, the focus of shock waves uses the pulverizing action of FSW, which is unacceptable for therapeutic influence upon internal body organs. Another disadvantage of the available equipment for SWT is that the diameter of the focus of a shock wave varies from 3 up to 12 mm that does not meet the size of the area of damage or pathological process and particularly organs' pathology. Thus, the available equipment does not essentially differ from lithotriptors, which are destined for extracorporeal lithotripsy, and hence, there is no essential difference either in cost or in maintenance charges.

NEW OPPORTUNITIES FOR THE METHOD OF SHOCK WAVE THERAPY

Experimental investigations in studying the biological action of shock waves upon organs and tissues have allowed establishing parameters of their stimulatory influence acting upon cell metabolism and reparative processes that has led to the development of a pilot sample of the shock wave device "Biostim" (Andriyanov, Yu.V., Garilevich, B.A., Cokhan, V.E., & Kudryavzev, Yu.V., 2000). Besides, the development of the lithotripter "Urat-P2" (Russia) has introduced modes of operation of the shock waves generator, which also have positive therapeutic effects. These devices use an electrohydraulic system for shock wave generation.

The first stage of the examination with regard to determining the parameters of stimulatory influence of shock waves has included experiments on animals. The influence of shock waves has been experimented upon the kidney at simulating acute and chronic pyelonephritis and wound surface.

The examination has found out that therapeutic action upon metabolic processes in tissues produces microsecond SWI varying from 3 up to 20 MPa in amplitude, and from 15^0 up to 60^0 in angulation of SWI front towards the time axis with a pulse duration at a half-height its amplitude varying from 0.6 up to 2.0 μsec. These processes are effective with pulse compression amplitude (i.e., positive phase of the pulse) less than 20 MPa and pulsing more than 500 pulses.

Experiments with a provoked skin wound have shown that the shock wave stimulation (SWS) made the wounds heal 5 to 7 days earlier than in the control group. Histological analysis of the healed wound has found out that the epidermis has been correctly formed having a normal position of the cell layers and well expressed dermal papillae. The control group of animals has shown a large area of the expansion of dense fibrous connective tissue, both in derma and in subcutaneous fatty tissues, as well as smaller integrity of smooth muscle fibers of derma.

Histological analysis of the kidney tissue after SWS has shown changes with evidence of metabolic processes activation. There has been present hyperemia of microcirculation blood vessels of the cortical and cerebral matter of the kidneys, mainly capillaries and arterioles. This specifies an increase in efficiency of extrarenal neurohumoral regulation of homeostasis due to the intensification of prostaglandin synthesis.

The experiments through simulation of acute pyelonephritis in animals have shown that 21 days after SWS the inflammatory infiltrates have been localized only in the submucous layer of the renal pelvis and have been of insignificant size. The control group of the experimental examination has shown massive leucocytic infiltrates present over the entire surface of the organ. The inflammatory processes have affected the cerebral substance by small peritubular infiltrates. Hence, a low-power shock wave influence acting upon the kidney at simulating acute pyelonephritis helps to reduce the inflammatory processes until their complete suppression. Besides, it has been found out that FSW with direct action upon microorganisms do not lead to the stimulation of their reproduction. On the contrary, the intensity of their reproduction reduces to a greater degree by direct influence of the focal area of a shock wave (Garilevich & Kudryavzev, 2003).

The experiment through simulation of chronic pyelonephritis (interstitial nephritis) and use of shock waves in a stimulation mode has proven that the kidney tissue has had a decrease in the amount of inflammatory infiltrates with prevailed macrophages and hystiocytes as well as activated fibroblasts in its cell structure. The amount of fibroblasts in the control group has been much lower. It has been found out that the sclerosis extent of the kidneys tissue in the control and experimental groups has been practically equal. Hence, low-power shock wave influence in case of chronic pyelonephritis has a sensible effect only on the suppress extent of the inflammatory reaction and does not have any essential action upon the extent of sclerotic changes in the kidney tissue (Kudryavzev, Garilevich, Kirpatovsky, & Olefir, 2003).

The results of the experimental examination of influence of microsecond electrical fields with amplitude varying between 1 and 7 kV/cm and shock waves with pressure up to 100 MPa upon permeability of the cell membranes of an ascitic tumor of mice in vitro have shown an intensification of the efficiency of irreparable pore formation under their synchronous action.

Under a simultaneous influence acting upon cell suspension of shock waves and electrical field of moderate intensity the survival rate of cells is 5 times less than under their nonsimultaneous action, and 10 times less in comparison with control parameters. Most sensitive to the influence of the investigated fields are cells in the phase of DNA synthesis, preparation for cell division, and in the phase of mitosis.

SWS of the area of a shinbones fracture (30 patients) has marked callus formation 9.6 ± 0.1 days earlier than in the group of patients where traditional methods of treatment have been used. In the experimental group of patients, the pain syndrome has been reduced 5 to 7 days in comparison with traditional methods of treatment. By the same term, the extent of movement in the contiguous joints has increased 10 to 15 degrees, and an improvement of regional blood flow has been marked. Stabilization of hemodynamics and microcirculation has promoted a restoration of the injured extremity function 8.7 ± 2.3 days earlier in 75% of the cases.

Taking into account positive results of the experimental and clinical examinations, we have used SWT method of therapy for treatment of the patients with chronic prostatitis (35 patients). SWT has been carried out using a lithotripter “Compact LGK” (Russia) with electromagnetic system of shock wave generation. The defocusing of shock waves has been constructively performed. The procedures were carried out every other day. A treatment course consisted of seven procedures. The control group included 15 patients who received a traditional physiotherapeutic treatment within a complex therapy.

All the patients of the experimental and control groups have shown a decrease in disease symptoms. In addition, the patients of the experimental group have felt a clinical improvement (inclusive according to the laboratory data) on the fifth day, on the average, while in the control group on the seventh day ($d<0.05$).

The control aftertreatment examination has statistically revealed a reliable improvement of uroflowmetry parameters that have been more expressed in the experimental group ($d<0.05$), as well as a decrease in disease symptoms evidence through a questionnaire survey using the NIH-CPSI scale ($d<0.05$).

Doppler color mapping has found a restoration of the symmetry of the prostate gland vascularization at 83.3% of the experimental group and at 80% of the control group. A transrectal ultrasonic examination in the grey scale mode has revealed a normalization of the prostate gland sonomorphology at 66.7% in the experimental group and at 60% in the control group. In addition, the most significant changes have been observed at the patients of the experimental group with fibrous changes in the prostate gland, who have shown after treatment a decrease in hyperechoic areas. The patients with dense calcifications have no improvement.

Thus, the preliminary results of use of SWT method on patients with chronic prostatitis have shown its efficiency that determines a perspective on the continuation of investigations. The best results can be obtained in the patients with fibrous changes in the prostate gland within lasting inflammatory processes, the treatment of which by traditional methods is insufficiently effective.

FUTURE TRENDS

Nowadays, there is a necessity to continue investigations oriented to the study of biological action of shock waves. It is necessary to optimize the parameters of shock waves, which can be used for extracorporeal fragmentation of calculi, stimulation of metabolic processes in organs and tissues, as well as for inhibitory action in case of malignant cell growth

With regard to extracorporeal shock wave lithotripsy, the problem has not been resolved by the present time, both concerning the quality of calculi fragmentation and injuring action of focused shock waves upon the kidney tissue. The achieved results of our investigations allow resolving this problem.

As regards the application of SWI stimulatory action, there is a factual opportunity to expand significantly the scope of application of the method in case of different diseases and pathological states.

The offered way to suppress tumors in the course of a complex treatment of patients with malignant neoplasms has prospects for development (Andriyanov et al., 1999).

The development and introduction in the clinical practice of a multibeam generator of shock waves (Andriyanov, Garilevich, Zotov, & Olefir, 2005) allow developing a multipurpose shock wave complex (for extracorporeal fragmentation of urinary and biliary calculi, shock wave stimulation of organs and tissues, as well as inhibitory action upon tumor cells).

CONCLUSION

On the basis of the data from articles and our own investigations, we have found that shock waves lead to an intensification of the reparative processes and metabolism stimulation in tissues and cells. SWT has found a wide application in orthopedic and traumatic surgery for treatment of patients with diseases of the locomotor system. It has been determined that the scope of application of the given method can be considerably extended, using stimulatory and inhibitory action of shock waves. Insufficient knowledge of biological action of shock waves upon organs and tissues and imperfection of the available equipment, which does not let shock wave pulses acting upon the internal organs, still impede.

The obtained theoretical, experimental, and clinical data allow continuing further investigations aiming at designing specialized equipment for SWT and expanding the scope of application of this perspective method.

REFERENCES

Andriyanov, Yu.V. et al. (1999). *The method of tumor depression*. Patent RU № 2127615. Bul. N 8. Moscow: FIPS.

Andriyanov, Yu.V., Garilevich, B.A., Cokhan, V.E., & Kudryavzev, Yu.V. (2000*). The device for stimulation of tissues metabolism by shock wave impulses.* Patent RU № 2151559. Moscow: FIPS.

Andriyanov, Yu.V., Garilevich, B.A., Zotov, V.N., & Olefir, Yu.V. (2005). *The multi-bean generator of focused shock waves.* Patent RU № 2247539.

Eisenberger, F., & Miller, K. (1987). *Urologische Steintherapie* (ESWL and Endourologit). Stutgard, New York: Georg Tieme Vtrlang (p. 999).

Garilevich, B.A., & Kudryavzev, Yu.V. (2003). The influence of shock wave impulses on microorganism reproduction intensity. In *Proceedings of the Plenum of the governing board of Russian society of urologists,* Sochi (pp. 108-109).

Garilevich, B.A., Kudryavzev, Yu.V., & Dzeranov, N.K. (2003). Stages of scientific elaboration in the field of shock waves application. In *Proceedings of the Plenum of the Governing Board of Russian Society of Urologists,* Sochi (pp. 106-108).

Garilevich, B.A., Zakharov, V.N., Kudryavzev, Yu.V., & Kirpatovsky V.I. (1997). *The method of tissues me-*

tabolism stimulation. Patent RU № 2076641. Bul. No 10. Moscow: FIPS.

Gerdesmeyer, L., von Eiff, C., Horn, C., Henne, M., Roessner, M., Diehl, P., et al. (2005). Antibacterial effects of extracorporeal shock waves. *Ultrasound in Medicine and Biology, 18*(1), 115-119.

Haupt, G. (1997). Use of extracorporeal shock waves in the treatment of pseudarthrosis, tendopathy and other orthopaedic diseases. *J. Urol., 158*(7), 4-11.

Kohri, K. et al. (1990). Effect of high energy schock waves on tumor cells. *Urol. Res., 18*(2), 101-105.

Kudryavzev, Yu.V., Garilevich, B.A., Kirpatovsky, V.I., & Olefir, Yu.V. (2003). The influence of low-energy shock wave impulses on a kidney tissue in a model of chronic pyelonephritis. In *Proceedings of the Plenum of the Governing Board of Russian Society of Urologists*, Sochi (pp. 179-180).

Mastikhin, I.V., Nicolin, V.P., Teslenko, V.S., et al. (1995). The increasing of sensitivity of tumor cells for cyclophosphane as result of shock wave effect. *Doklady AN RF, 342*(2), 262-264.

Mironov, S.P., Vasiljev, D.O., & Burmakova, G.M. (1999). Extracorporeal shock wave therapy of chronic degenerative inflammatory orthopaedic diseases in sportsmen and ballet-dancers. *Vestn. Sport. meditsiny Rossii*, (1), 2-4.

Nishida, T., Shimokawa, H., Oi, K., Tatewaki, H., Uwatoku, T., Abe, K., et al. (2004). Extracorporeal cardiac shock wave therapy markedly ameliorates ischemia-induced myocardial dysfunction in pigs in vivo. *Circulation, 110*, 3055-3061.

U.S. Patent No. 3735755 (1975).

U.S. Patent No. 4441486 (1986).

W.O. Patent No. 86/ 05104 (1986).

Siebert, W., & Bush, M. (1997). *Extracorporeal shock waves in orthopaedics*. Berlin, Heidelberg, New York: Spring.

Valchanov, V.D., & Michailov, P. (1991). High energy shock waves in the treatment of delayed and non-union of fractures. *International Orthopaedics, 15*, 181-184.

Zubkov, A.D., Garilevich, B.A., & Olefir, Yu.V. (2005). Experimental study of low-energy shock wave impulses effect on metabolic and reparative tissue processes. *Voenno-meditzinsky zhurnal, CCCXXYI*(9), 59.

KEY TERMS

Focus of the Shock Wave: Area with high-amplitude pressure up to 300 – 1000 atm.

Focused Shock Waves: Shock waves are focused by ellipse-formed reflector into definite area.

Metabolism Stimulation: The metabolism is improved on the level of organs and cells.

Shock Waves Parameters: Pressure amplitude and duration of positive and negative shock wave impulse phases.

Shock Wave Therapy: The application of shock waves' curative effect.

Reducing Patient Delays in a Day Surgery Unit of a Hospital

Victoria Hanna
University of Melbourne, Australia

Kannan Sethuraman
Melbourne Business School, Australia

INTRODUCTION

In this article, we illustrate the relevance of queuing theory principles to the healthcare sector through a case study of a day surgery unit in a hospital in Melbourne, Australia. The hospital has an acknowledged problem: patients are spending longer-than-anticipated periods of time in the day surgery unit, and they are facing excessive waiting times at all stages of their healthcare delivery process. The hospital is operated on a not-for-profit basis, and the executive board members are keen to understand the root causes of the problem, so they can direct their investment in the system to improve its responsiveness.

BACKGROUND

To reduce the time a patient spends waiting in the hospital system, a hospital needs to understand the implications of queuing and the theories that explain the causes and costs of congestion. Little's Law (Little, 1961) states that the average number of patients in a stable system (over some time interval) will equal the product of their average arrival rate and their average time in the system. In manufacturing situations, this law represents the critical link between the three key internal process performance measures: cycle time, work-in-progress (WIP) inventory, and manufacturing lead time (MLT).

- Cycle time characterizes the average time between completions of successive units. In the healthcare context, for instance, the cycle time could represent the average time between completions of successive procedures in a surgical unit.
- WIP inventory represents the number of units contained in the process at any given time. Patients-in-progress (PIP) will be the equivalent measure in the healthcare context, representing the number of patients who are currently being served, and who are waiting for service in the system.
- Manufacturing lead time stands for the total time a unit spends in the manufacturing process. In the healthcare context, the patient resident time (PRT) would be its equivalent. And it represents the total average time a patient spends in the hospital, from the time of his arrival, to his discharge time. This would include all the value-added time and the nonvalue-added time, such as waiting that the patient experiences in the system.

Using the average number of patients residing in the system at any given time, and the surgery throughput capability of the system, Little's Law helps us predict the average time a patient would reside in the system (PRT). In other words, the larger the number of patients waiting for treatment, the longer will be the wait for an arriving patient. An immediate implication of Little's Law is that, for a given level of throughput at any hospital, the only way to reduce the PRT is to reduce the number of patients in the system. For example, assuming patients arrive at the rate of 10 per hour, and stay an average of one hour. This means we should find the average number of patients in the system at any time to be 10. If the arrival rate rises to 20 per hour, the unit must either be prepared to host an average of 20 occupants, or must reduce the time each patient spends in the system to 0.5 hour. Naturally surgery cannot be sped up, so the unit must look to other components of the patients' stay to reduce their time in the system.

This relationship, however, is not obvious to hospital administrators or doctors. The system expects people to wait, and hence provides infrastructure to deal with patients waiting in the system, resulting in escalating costs and poor responsiveness. Long lead times are still accepted by patients, as they have no alternative. What the service providers are missing is an opportunity to simultaneously improve both patient responsiveness and cost effectiveness. Too many patients in the system can be costly in a number of ways. You need more room to hold them, more resources to engage them, monitor, and progress them through the system (hospital), and the excess number of patients in the system acts as a buffer, hiding the linkages between processes. It is critical to observe that the time spent delivering patient care should never be sped up (for example, time for anaesthesia to start working, treatment of wounds, recovery time, and so on), as it compromises quality of care to achieve responsiveness. The focus has to be on simultaneously achieving responsiveness and quality of care through the elimination or minimisation of nonvalue-added activities, such as waiting for service.

How Variability Increases the Time Patients Spend in the System

Variability robs a system of capacity. As a consequence, utilization is reduced, and this affects the time a patient spends in the system. The *Pollaczek-Khintchine* formula (Heyman & Sobel, 1982) provides the fundamental relationship between capacity utilization, variability, and inventory, and is represented by:

$$L = K \times \left(\frac{\rho}{1-\rho}\right) \times \left(V_a + V_s\right)$$

where K is a constant, ρ represents the capacity utilization, V_a characterizes the variability in arrivals, V_s characterizes the variability in the service process, and L stands for the queue length.

If patients undergoing treatment are a form of inventory (analogous to work in progress), the above relationship has important consequences for a hospital trying to reduce patient delays.

- In steady state, if a hospital system experiences higher levels of variability (either in the patient arrival process, or in the patient care delivery process, or both), it will essentially result in an increased patient queue length. From Little's Law, we can infer that increased patient queue length will lower responsiveness, or in other words, increase average patient resident time.
- In highly unpredictable environments (such as an emergency department), the patient queue length and the average patient resident time tend to increase exponentially with increased capacity utilization. If better information cannot be obtained to lower the level of variability in the patient arrival process, the system will need to operate at lower levels of utilization to achieve the target service levels. That is, they will need buffer capacity. Attempts to fully utilize capacity will result in extremely poor responsiveness—a situation unlikely to be acceptable to patients or healthcare providers.

Litvak and Long (2000) discuss how healthcare systems exhibit variability, and categorise them into three broad types: (i) *clinical variability,* which represents the differences in the patients' degree of illness, choice of treatment alternatives, and responses; (ii) *flow variability,* which captures the randomness in the arrival patterns of these patients; and (iii) *professional variability,* that describes the variability in the ability of the medical practitioners and healthcare delivery systems to provide treatment. The presence of clinical, flow, and professional variability results in a significant increase in complexity, and adds cost to healthcare systems. In addition to these natural variabilities, they also identify an artificial variability, which is an artefact of dysfunctional management and policies. They conclude that eliminating artificial variability in healthcare systems holds the maximum potential for reducing waste and improving responsiveness.

These theoretical underpinnings helped us to explain why the day surgery unit in our case was experiencing excessive wait times for its patients, and our analysis of the situation focused on the following:

1. **System design:** If the patient care delivery process is poorly designed (for example, redundant activities, activities occurring in sequence instead of in parallel, and so on), then the total time it takes for a patient to progress through the system will be longer than ideal.
2. **Waiting times at key stages of the process:** Even when a system is well designed, the wait times could be poor, due to constrained resources acting as bottlenecks. Since all systems exhibit some variation, queuing theory provides key insights

into why wait times deteriorate sharply in highly utilized systems.

3. **Variability:** Even well designed systems, without any limiting resources acting as bottlenecks, can still display poor responsiveness. This can be attributed to the level of variability in the system.

DATA COLLECTION

The case study of the day surgery unit was developed from direct observation of its processes, interviews with hospital staff, and review of secondary data. A total of 17 detailed interviews were carried out over a period of six weeks. The interviewees included the doctors, anaesthetists, nurses, technicians, middle managers, and the hospital executive team. Secondary data sources included a survey of surgeon satisfaction with the facilities, a survey of patient satisfaction focusing on the patient's experience at the day surgery unit and the extent to which it met their expectations (both of these surveys were deployed by hospital staff immediately prior to the involvement of the authors), internal project reports describing improvement activities undertaken in the previous five years, key performance indicator results, value and mission statements, and published material targeted at patients.

THE CASE

The day surgery unit handles only elective surgeries, and annually processes about 10,000 overnight admissions and 7,400 day surgeries. They offer a range of specialties, including orthopaedics, neurosurgery, urology, plastics, breast surgery, and dental surgery. The hospital provides facilities, equipment, and nursing staff. The hospital facilities consist of nine operating rooms (OR) with state-of-the-art equipment, and hospital wards for in-patients. OR equipment may either be owned by the hospital or be procured from outside on a loan/hire basis. The equipment that is owned by the hospital will usually be shared across theatres. Regulatory requirements specify threshold levels for nurse-patient ratios, and inability to provide adequate number of qualified nurses to meet the ratio requirements can result in "unused" beds and consequent underutilization of available capacity.

Scheduling Surgeries

Prior to interacting with the hospital, the patient will consult with his or her primary care physician. Upon receiving a referral from the general practitioner to visit a medical specialist, the patient will be assessed by the specialist. Once the need for surgery is established, the patient's name will be added to the surgeon's list of patients requiring surgery/hospitalization. Typically, specialist doctors have privileges at several hospitals in the area, and the choice of hospital for surgery is determined by a number of factors that include cost, patient preference, case complexity, wait involved, facilities, and other services provided by the hospital. The case hospital competes on location, range of services offered, and its reputation for quality and patient care. It is not uncommon for elective surgeries to be scheduled several weeks (even a few months) in advance. The day surgery unit receives information of surgeon lists as late as 24 hours prior to surgery. The surgeon's operating list is dynamic, with additions possible just a few hours ahead of surgery session start time. The hospital is expected to be responsive, and be able to provide the requisite support of staff, equipment, and supplies.

Similar to other hospitals, manually derived annual theatre plans form the basis of the hospital's resource allocation. The theatre plan involves assignment of theatre sessions to consultant surgeons, and essentially defines the demand for hospital's services. A half-day slot is considered the basic unit for this purpose, and surgeons expect hospitals to provide complete flexibility in organizing the activities within their sessions. The assignment of a theatre session to a surgeon commits the hospital for providing the necessary staff and equipment for surgery and postoperative care, and hence, the hospital workload is a direct consequence of the theatre plan.

Typical Patient Flow

When the patient arrives, they first go to the hospital reception. At this stage, they hand over any test results or medical notes they have brought with them, and complete the necessary information regarding private health insurance. They will also be advised of any additional payment necessary. Patients are seen on a first-come first-serve basis, rather than in the order their name appears on the surgeon's list. All of the patients

R

having surgery in the morning session (across all of the nine ORs) are requested to arrive between 6:45 and 7 a.m. There can be between 30–50 people attempting to access the reception desk at this time. Once the business office (behind reception) has completed all the necessary paperwork, the patient proceeds to the Surgical Admission Unit (SAU) reception. At the SAU reception, the patient's medical records are checked again, and any discrepancies are followed up. The patient is requested to change their clothes and proceed to the SAU waiting room. Most patients are accompanied by family members, and there is a small waiting room area adjoining SAU where they can stay. Unless they are accompanying a child, the visitors are not permitted to enter the SAU.

All patients are required to complete nurse education before being transferred to OR, and the majority of cases also require preoperative assessments by their respective anaesthetists. Nurse education involves meeting with one of the SAU nurses who will explain how the day will proceed. This could be done at the same time as admission, but when there are large numbers of patients waiting, it is often delayed to a later stage. Due to the congestion in SAU, these meetings occur on an ad hoc basis. Both anaesthetists and nurses can be observed searching for their patients, trying to ensure the first patients on the surgeons' lists are ready to enter OR on time. The anaesthetists are under even more time duress, because they want to see all their patients in a given session before surgery commences on the first patient. The reason for this practice is that the anaesthetists want to be physically present in the theatre for the complete duration of surgery, and they are also under pressure from surgeons to ensure fast turnaround between surgical cases. Another contributing factor is the physical layout of the theatres, which are not in proximity to the preoperative assessment rooms.

Once the surgery is complete, the patient is transferred to a recovery ward. Once fully recovered, he walks back to SAU and gets dressed. If no further payment is required, he is then discharged via SAU. If additional payment is required, he is directed back to the hospital reception, whereupon he will make final payment to the hospital, and is discharged by the business office.

The patients for the afternoon session all arrive between 10:30 a.m. and 11 a.m., repeating the chaos of the morning. Table 1 illustrates how long patients are waiting at various stages of their day. Some patients move rapidly through the system, but others can spend long periods of time waiting to enter both the SAU and the OR. On an average, a patient resides more than six hours in the system, where the actual value-added time is less than an hour in duration.

DISCUSSION

Through a detailed examination of the patient flow, several factors that contribute to delays were identified, and are presented in Table 2.

PROPOSED ACTIONS

The researchers, executive board, surgeons, anaesthetists, and nurses brainstormed various approaches that may help the situation and solve the problems identified in Table 2. Based on these discussions, the

Table 1. Typical patient wait times for day surgery cases

Description	Min Time (minutes)	Max Time (minutes)	Average (minutes)
Patient waiting at reception	1	54	13
Patient waiting for admission into the Surgical Unit	1	154	25
Pre-operative interview with anesthetists	1	18	4
Surgery duration	10	180	30
Time from entering SAU to entering OR	11	273	128
Total time in system	64	584	374

following priorities were developed for the hospital to focus on.

System Design

It was decided that it was not possible to make any changes to the physical design of the system, but was possible to modify how patients interacted with the system on arrival. It was agreed that when patients arrived at the hospital reception, their details would be noted, and they would be immediately asked to take a seat, rather than be admitted. The admittance procedures would be managed by calling patients in order of their position on the surgical list. The reception staff was briefed on the new protocol, and posters were put up to advise patients that they would not necessarily be called to the office in order of their arrival. The aim of this proposal was to reduce the number of times commencement of surgery was delayed, because the first patient on the list was not ready.

The second change focused on moderating the volume of patients in the system. Staged admission with risk stratification was proposed as a method to eliminate bulky arrivals. The major contributor to long patient lead time was the requirement by anaesthetists and surgeons for all patients in a given session to arrive at the same time. A staged admission process with risk stratification where patients did not arrive en masse was proposed. The practice is already established at several public hospitals. It requires the provision of complete patient information at the time of scheduling a surgery. In this scheme, patients are typically stratified into risk categories (low, high, very high) on the basis of institutional practices and management guidelines and known risk factors. All high-risk patients will be required to arrive at the hospital at the start of the day to ensure appropriate evaluation can be conducted prior to the start of their surgical procedures. This proposal is aimed at lowering the peak load of patient arrivals, and thereby reducing the pressure on resources, and considerably reducing the wait patient's experience.

Reducing Artificial Variability

The hospital also recognised the need to minimize all dysfunctional management practices that artificially escalate the impact of existing normal variability. This will require better information exchanges between doctors and hospital staff, as well as improved coordination across all stages within the hospital. Sharing accurate, real-time patient information to all parties would also aid in lowering uncertainty. Basic protocols for updating and transferring information were proposed, and responsibility for ensuring the consistent execution of these protocols was incorporated in the role of the unit's nurse-manager.

Table 2. Leading factors contributing to the poor performance of the day surgery unit

System Design	*Bulky arrivals*	The large numbers of patients arriving at once leads to bottlenecks, both at admission, and preoperative assessment.
	Physical design of facilities	The size of the unit does not accommodate the volume of patients being processed at the beginning of each theatre session. In addition, the ORs are separated by geographical distance, and insufficient time is given for the transfer and preparation of shared resources between theatres.
Waiting Times	*Resource bottlenecks*	Insufficient capacity of ward beds, recovery area, orderlies, and so on can result in blockage and starving of ORs. For example, if the recovery ward is full, the ORs cannot release their patients, and delays start to build.
Variability	*Lack of coordination and information sharing*	Missing information/test results, insufficient advance planning, due to lack of information and lack of prioritization, all lead to delays in patient flow. Information is updated in a decentralised manner, causing inconsistencies and duplication of effort. Data is also not updated in a dynamic fashion to reflect accurate information.

OUTCOMES

These simple changes to administrative procedures were helpful, and had an immediate impact. The staged admissions process was piloted with one surgeon and associated anaesthetists with positive results. The board was pleased with the improved performance, and has committed to examining all processes, not just those inside the day surgery unit, and has engaged outside consultants to coach the employees through a large-scale evaluation of their systems.

CONCLUDING REMARKS AND FUTURE TRENDS

This research has broader implications for strategic decision making and improved alignment between hospital priorities and their operational policies. Drawing on the insights of Lovejoy (1998) who noted the substitutability of inventory, capacity, and information (information acts as a surrogate for variability), to achieve a given service level, a hospital can achieve their desired responsiveness (expressed in terms of acceptable wait times for treatment) in one of two ways: (i) it could either hold surplus capacity when it is unacceptable to keep patients waiting or it is difficult to predict the arrival of patients or their needs, or (ii) it could invest in collecting enough in-depth and real-time data that would aid in lowering the variability in demand/services, and thereby using its available capacity more effectively without undue wait for its patients. Alternatively, they could relax the responsiveness requirement if the patients in their focused segment are not particularly time-sensitive, and are willing to wait for their treatment. In this situation, the hospital can continue to operate with high-capacity utilization and higher levels of variability in its demand/services.

The case also highlights how artificial variability adds complexity and costs to the healthcare system, and in the absence of perfect coordination and exchange of information between the various links in the healthcare value chain, the hospitals are forced to keep surplus capacity to be able to provide flexible and responsive services to their surgeons, resulting in higher costs. Also, the patients spend a much longer time in the system when there are capacity constraints within the system. It is essential hospitals understand the trade-offs that exist between (i) buffer capacity, (ii) collecting good quality information to lower variability in demand, (iii) standardisation of processes to lower variability in service provision, and (iv) patient resident time.

Through this article, we have also shown that operations management principles such as queuing theory, originating in the manufacturing sector, have far reaching implications for the healthcare sector. However, to make this application more readily acceptable to the healthcare audience, it is critical to translate the core components of operations management principles into a form that directly applies to this context. For example, if a hospital plans to use lean principles to lower the level of waste in their system, it would be first necessary to characterize what constitutes waste in a healthcare context. A taxonomy of waste as it applies to this sector would enable both healthcare practitioners and managers to better direct their efforts at targeting waste and improving the patient experience.

REFERENCES

Heyman, D., & Sobel, M. (1982). *Stochastic models in operations research, volume I*. New York: McGraw Hill.

Little, J. D. C. (1961). A proof of the queueing formula L = λ W. *Operations Research, 9*, 383-387.

Litvak, E., & Long, M. (2000). Cost and quality under managed care: Irreconcilable differences? *The American Journal of Managed Care, 6*(3), 305–312.

Lovejoy, W. (1998). Integrated operations: A proposal for operations management teaching and research. *Production and Operations Management, 7*, 106–124.

KEY TERMS

Artificial Variability: The direct result of the poor management of the processes used to provide care. This type of variability (dysfunctional management and policies) can be reduced through improved operational policies and system design.

Natural Variability: A source of great waste in the health care delivery system is excessive variability in the processes used to provide care. Natural variability is largely outside the control of a hospital, and it includes "clinical" variability (patients differ in the type and severity of their diseases, and patients with similar ailments respond differently to treatment), “patient demand” variability (patients arrive for treatment randomly over time), and “professional” variability (different surgeons treat similar patients in different ways), which has given rise to the development of approaches like practice guidelines and clinical pathways.

Patient Resident Time (PRT): The amount of time a patient spends in the healthcare system, typically measured from the moment they enter the system, until the time of their medical discharge.

R

Reforming Nursing with Information Systems and Technology

Chon Abraham
College of William and Mary, USA

INTRODUCTION

Much of healthcare improvement via technology initiatives addresses gaining physician by-in (Reinertsen, Pugh & Bisognano, 2005) and does not adequately address engaging nurses, despite the fact that nurses serve as the front-line caregivers and are a primary user group (Wiley-Patton & Malloy, 2004). However, the tide is changing, and visibility of nurses as information gatherers and processors in the patient care process is increasing (Romano, 2006). Nurses perform the majority of the data-oriented tasks involved in patient care and would benefit most from having access to information at the point of care (Bove, 2006). RADM Romano, Chief Professional Officer of Nursing and advisor to the U.S. Surgeon General concerning public health, recently addressed the American Informatics Nursing Informatics Association and stated, "This is the year of the nurse. The technologies that have the means to improve the efficiencies in patient care are in the hands of the nurses." Nurses need to embrace technology in everyday work or continue suffering the consequences of antiquated methods of computing that take us away from where we work—at the point of care (Abbott, 2006).

Healthcare can benefit greatly from use of a diverse set of technologies as facilitators of change to improve quality and safety of patient care by decreasing errors made because of the lack of faster, more comprehensive, and more accessible patient documentation at the point of care (IOM, 2000). Healthcare institutions abroad share similar sentiments according to international healthcare reports by the International Council of Nurses (Buchan & Calman, 2004). In light of these concerns and a severe U.S. nursing shortage (i.e., an estimated 400,000 shortage by 2020) (Bass, 2002), institutions are beginning to consider employing point-of-care IT as a means to promote patient safety, decrease medical errors, and improve working conditions for overtaxed nurses (TelecomWeb, 2005). This chapter provides an overview of nursing reform and novel ubiquitous computing integrating voice, hands-free, and mobile devices being researched or used to make nursing more efficient and promote the following:

- A decrease in laborious documentation aspects (i.e., decrease problems with access to information at the point of care, written errors or legibility issues precluding comprehension of medical regimens).
- An increase in recruitment and retention rates.
- Improvement in communication in which nurses are required to consolidate and process information during care.
- Promotion of involvement in systems analysis and design of information systems to increase the likelihood of technology acceptance.
- Baseline and continuing education in both professional training and on-the-job training to allay technology aversion and build computer self-efficacy.
- Development of standards for electronic documentation of patient interaction and processes for nursing that can be codified.
- Identification of role changes in the nursing community because of the use of information systems and technologies.
- More rigorous research concerning information research in the nursing community.

BACKGROUND

This chapter provides an overview of tactics to reform nursing sponsored by leading nursing healthcare information systems research in professional organizations such as the American Nurses Informatics Association (ANIA), Association of Medical Informatics (AMIA), and the International Council on Nursing (ICN).

The sociotechnical systems framework (STS) (Bostrom & Heinen, 1977) will be employed to categorize tasks, technologies, people involved, roles, and structure of aspects of the reform. STS emphasizes workplace interactions with various technologies and is espoused as a realistic view of organizations and a way to change them (Bostrom & Heinen, 1977). STS concerns the social system that is comprised of contributions of people (i.e., attitudes, skills, and values), the relationships among people and their roles in the organization, reward systems, and authority structures (Bostrom & Heinen, 1977). STS is also an intervention strategy in which technology implementations intervene within a work system to improve task accomplishment, productivity, and work quality of life, and to prompt supportive organizational structural changes. Innovative technologies may improve task efficiency and task effectiveness by automating or reengineering antiquated/manual processes, changing people's roles, and making organizational structures (Bostrom & Heinen, 1977). A graphical depiction of STS is displayed in Figure 1.

DESCRIBING NURSING REFORM USING AN STS FRAMEWORK

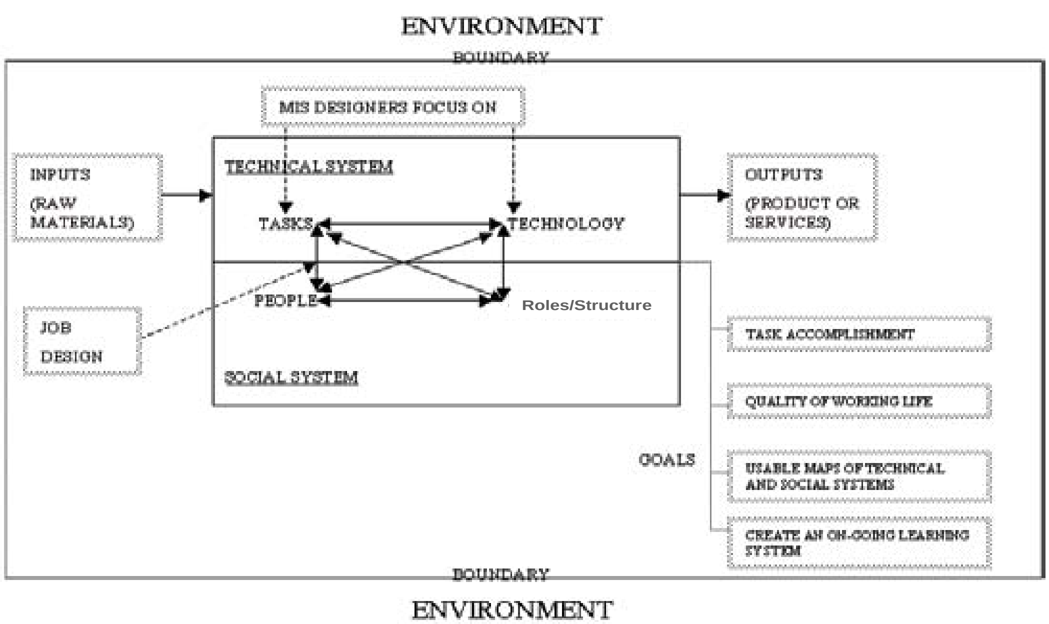
STS is used as a frame to describe elements addressed in the nursing reform initiative, such as tasks, technology, people involved, changes in roles, and structure in conjunction with their interplay, outputs, and goals.

Task

The tasks in this reform are indicative of the pervasive problems plaguing nursing and healthcare in general. There is a critical nursing shortage crisis in the United States alone in which more than 90,000 nursing position vacancies were reported in 2004, and nearly 400,000 are expected by 2020, which has grave impacts on quality of care (JCAHO, 2004). Similar statistics reported by the International Council of Nursing (ICN) also highlights the great supply and demand disparities. Much of this is attributed to the problems recruiting new nurses and retaining seasoned ones. Nursing is an aging workforce in which the following is reported:

- One in five registered nurses plan to leave the profession.
- 81% say morale is extremely low.
- 64% do not have enough time to spend with the patient.
- 60% note the manual paperwork burden to document information about patient care tasks is "a necessary evil."

Figure 1. STS (Bostrom & Heinen, 1977)



The last percentage indicates the manual paperwork burden, which is one problem that can be directly addressed via application of novel information systems (IS) and information technology (IT). Information-oriented patient care tasks for which nurses are primarily responsible entail the following:

- **Triage:** The process of taking vital statistics, documenting impending problems, and making initial assessments to categorize criticality of the emergent condition.
- **Charting:** The process of monitoring the patient and recording vital statistics, effectiveness of medical intervention, and overall patient wellness, which is an ongoing process across nurse shifts.
- **Throughput flow:** The process of managing when, where, and how patients are moved about in a patient care system across units within the hospital for labs, procedures, and so forth.
- **Medication administration:** The process of validating medication regimens with prescriptions, delivering the medication, and documenting the effectiveness.

Virtually 98,000 people die each year due to medical errors in hospitals (Institute of Medicine, 2000), and many of these are likely due to antiquated methods of information management and delivery during these fundamental patient care tasks. These errors can also be attributed to the inability of personnel relying on the information documented by nurses during these tasks to actually access it in a timely fashion. In essence, not having access to pertinent information when and where needed cripples the abilities of healthcare professionals such as nurses from validating medical information and diagnosing properly. So the primary task for reform becomes automating manual charting and equipping nurses, who are highly mobile in their work environment and move from patient to patient, with devices that enable easy and unobtrusive means of accessing and documenting needed information during patient care.

Technology

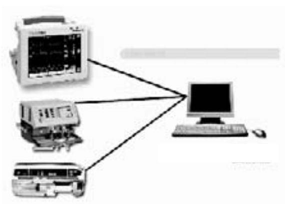
The reform in this industry recognizes that the paperwork burden is one where technology can be used as a facilitator of change or as part of the intervention strategy to promote change. Therefore, the goal is to promote productivity in which nurses are able to decrease errors associated with manual charting and reduce charting time via easy non-time-consuming data input and output capabilities. These improvements enable a nurse to spend more time with patients, which is a stated desire among nurses (Abraham, Watson, Boudreau & Goodhue, 2004). AMIA and ANIA encourage partnerships among vendors, academic researchers, and nurses in practice to discern technologies that are vehicles for promoting their workflow efficiencies. Table 1 provides an overview of these technology categories.

One aspect of technology that is specific to the healthcare industry and not akin to the average mobile business environment is the need to protect patients and caregivers from the transfer of infectious diseases or bacteria introduced by mobile devices. The typical mobile device is assigned to one nurse who carries it throughout his or her shift. Therefore, the nurse uses the same device/technologies across every medical intervention in which the likelihood of bacterial transfer is high. Thus, antimicrobial protective coatings or materials on healthcare mobile devices are of utmost importance in order to avoid transfer of bacteria from one patient environment to another, or from one nurse to another as the devices turn over during a shift change.

People

Nursing in general is a highly caring and nurturing profession (Patistea, 1999). Nurses have chosen their profession based on their individual desires to care for others and promote the sanctity of human life. Nurses, for example, who are satisfied in their jobs are characterized by high levels of affective empathy (i.e., the ability to be moved by other people's feelings) and service-mindedness (i.e., a strong tendency to let other people be central and dominant) (Sand, 2003). In addition, they display a high degree of sensitivity to nonverbal communication and a strong need for affirmation, support, and dominance (Sand, 2003). The traits prioritize patient care tasks for servicing nurses. For example if a patient is exhibiting physiological problems associated with anxiety or is in pain, the primary task becomes physical care of the patient. Additionally, since nurses do display a strong need for affirmation from their patients, peers, and superiors, the technologies need not interfere with the desire to

Table 1. Technology overview

Technology Category	Task Orientation	Purpose	Advantage
Bio-medical device integration 	Triage and Charting	Automatically capture patient vital statistics and other physiological metrics that otherwise would be taken manually such as blood pressure, oxygen level, etc through multiple point-of-care devices and provide data in a format that is easily integrated into the electronic medical record.	Decreases likelihood of human error or misreadings. Since the devices can be set to periodically take the measurements, they decrease the time required for scheduling to come back to the patient to take the measurements (i.e., promotes efficient use of personnel time).
Internal Communication 	Charting	Simplify staff assignments as for staff turnover intermittently during shift changes, provide internal communication that displays staff contact information, streamline patient bed turnover to notify housekeeping automatically.	Decrease administrative burden of assigning nurses to patients and provide access to nurses while on duty so that they can be more responsive.
Intra-hospital Patient Transport 	Throughput Flow	Provide a means to prioritize and schedule transport of patients with instructions and provides access to patient location information (i.e., the unit and time allocation for patients allotted transport time).	Eliminates the need for dedicated transport dispatch personnel, eliminates manual creation of transport schedules, enables on-time pick-ups to promote throughput flow efficiency.
Medication Administration 	Medication Administration	Provide access to patient medication regimen prescribed by the physician and validated by the pharmacist via a mobile medication workstation. Workstation is equipped with temperature controlled drawers, separated baskets to store medications in transport to the point of care and barcode scanners to scan medication labels and patient ID bands to automatically validate data with the patient medication information. Ensures that correct meds are being giving to the correct patient.	Decreases likelihood of errors in administering incorrect medications or dosages because of automatic validations and ability to have access to real-time information to be aware of medication changes. Saves time by eliminating the need to create meds list at the beginning of each meds pass that are not real-time . These static manually created lists lead to error if medications change unbeknownst to the administering nurse. Storage capacity minimizes the need for nurses to make multiple trips to the medication cabinet to obtain meds during the meds pass because the nurse can load virtually all meds needed in the medication workstation compartment at one time. Decreases the likelihood of spoilage or misplacement of medications because they are held in a temperature controlled compartments for each patients' mix of meds.

continued on following page

Table 1. continued

Technology Category	Task Orientation	Purpose	Advantage
Mobile Computing 	Triage and Charting	Provide access to medical information systems via untethered computers mounted on mobile workstations that nurses can maneuver with them as they move from patient to patient. Most likely a single workstation is provided to each nurse on shift who is servicing patients.	Allays problems associated with having to record information on a manual chart or rely on memory until the nurse can locate a computer to electronically document.
Wearable Technologies 	Triage and Charting	Less obtrusive technology geared for mobile workers via multimodal user interfaces for which the information system is operated via text, speech, touch, and gesture.	Keeps hands free to work with patients, enables easier maneuverability within workspace.

establish mutual trust with the patient and the need for nurses to exhibit confidence in use of the technology, especially in the presence of the patient.

The clinical nursing population can be divided into two demographics: A and B. Demographic A is an average age of 40, has work experience of 15 or more years, lacks professional computer training that is a systemic problem recognized by servicing IS personnel and nursing managers (Bass, 2002), and has low computer self-efficacy (Abraham et al., 2004). Demographic B is characterized by an average age of 25, has less than seven years experience, was trained under a "new" regimen that introduced him or her to PC-based documentation, and exhibits medium to high computer self-efficacy (Abraham et al., 2004). Both demographics use traditional PCs for other tasks, such as monitoring vital statistics or maintaining an apparatus for intravenous drug administration. However, information systems for electronic charting or mobile devices that provide access to these information systems are revered initially as foreign to this population (Abraham, forthcoming). The majority of nurses in Demographic A do not type effectively but regard narrative oriented data entry as problematic (Abraham et al., forthcoming). Additionally, the personality traits of nurses influence how they accept and use technology (Abraham et al., 2004).

In essence, intended utilization for mobile technologies is for the nurse to use at the point of care and, most likely, in the presence of the patient. Therefore, the technology should not impede nurses from being able to maintain eye contact nor interfere with the interaction with their patients. Additionally, healthcare workplaces are very tumultuous, hectic, and characterized by a great deal of uncertainty, which contributes to stressful work environments. Thus, the technologies need to be designed in a manner that are easy enough not to excite anxiety among the nurse user group that causes nurses to fumble with the technology, display a lack of confidence, or inhibit nurse/patient interaction (e.g., having to pay more attention to the computer because of ill-designed interfaces) (Abraham, forthcoming). Nurses feel that their lack of confidence concerning operating the technology is interpreted by patients as incompetence not only with the technology but with performing their required patient care tasks (Abraham et al., 2004). Not only does the technology need to support the way nurses work, but it also needs to support the way nurses think about servicing their patients. Also, adequate computer training, including how to most appropriately use the mobile technologies as vehicles to support the patient/nurse relationship, is of utmost importance.

Physical care of the patient will always overcome the need to document the intervention until the hands-on care has been performed. However, providing nurses with easy unobtrusive access to medical information when and where needed enables nurses to make better well-informed decisions.

Interplay Between Task, Technology, and People in Healthcare Environments

Mobile technologies make new dimensions salient that are driving novel interplay between the task, technology, and people in healthcare environments. For example, traits of the nurse concerning the need to attain affirmation or appear competent in front of the patient while operating the technology is novel since prior to mobile computing the act of operating the computer most likely was done away from the point of care (i.e., at the nurses' station). When computing using technologies that were tethered, nurses were less likely to exhibit as much anxiety about using the system in front of the patient (Abraham, forthcoming). The interplay among task, technology, and people in healthcare environments has become more pronounced because of changes in computing methods that enable access to information when and where needed via wireless technologies.

Nursing reform addresses initiatives in the development of technologies not only to better support the way nurses actually work, but also to develop information systems based on a unified nursing language known as the International Classification for Nursing Practice (ICNP). It is an arduous task to implement compositional terminology for the nursing practice since nurses affectionately note that each individual "nurses" differently (i.e., they label common patient care functions differently) (ICN, 2006). The goal of the ICNP is to provide a global infrastructure informing healthcare practice and policy to improve patient care via codifying and retaining tacit, experiential knowledge to inform the practice and promote easier cross-training, especially in light of the aging nurse workforce and the problems with recruitment and retention (ICN, 2006). In light of these globally recognized nursing shortages, persons with either formalized nursing training or interest in the field from foreign countries have been recruited and trained to fill many vacant positions (Bass, 2002). ICNP hopes to streamline the learning curve in order for these foreign nurses to become acclimated to the nursing profession.

Additionally, as people relocate, which is increasingly the case, their patient information should move or be automatically accessible despite the physical location of the patient (Council on Systemic Interoperability, 2006). However, the limiting factor is the lack of systemic interoperability within the United States and globally to make this idea a reality. ICNP focus is to ensure that at least care providers speak a common language, which is documented in the patient's medical record that patients transport when they relocate or acquire services from a multitude of different healthcare providers. Therefore, ICNP and systemic interoperability projects are addressing the interplay between fulfilling the task with suitable information systems and technologies to support the people who are both task doers and task recipients in this field in order to promote better quality of care—the ultimate goal.

Roles and Structure

These novel technologies and initiatives in improving the communication among nurses and about what they actually provide as value-added effort helps to better define and highlight the crucial role of nurses. By virtue of these technologies being placed in the hands of nurses, it is indicative of the technology being a facilitator of change to realize the role of the nurse as instrumental to the patient care process since they do perform the bulk of documentation for any medical intervention. Contributing to realization of the criticality of their role is the nurse's ability to recognize in an easier and more timely fashion the potential errors in the prescribed regimen, such as medication prescriptions that may have been prescribed erroneously or entered into the system mistakenly by a physician but validated by the pharmacist. In this sense, having access to IT and IS empowers nurses to at least bring to light problems they discern regarding the prescribed regimen for a patient. Typically, nurses that are assigned to in-hospital patients interface with their patients more frequently than any other medical personnel; they are very familiar with their patient's regimen and have the astuteness to recognize when there may be a potential error in the prescribed regimen. For example, the servicing nurse may discern allergic reactions to the prescribed medication or an ineffectiveness of the medication requiring a physician's order to be clarified or modified. Having access to information coupled with efficient means of communicating with other servicing medical personnel such as a patient's physician allows the nurse to discern the validity of the regimen or prescription in question. In essence, having access to information when and where needed and having the ability to communicate effectively because of the aforesaid novel technologies enables the nurse to appear more competent and thus gain more respect for his or her level of knowledge and critical role in the ability of the patient to be serviced properly.

As far as structural impacts, nurses contend that they have been somewhat neglected and overlooked as principal components in the care process, which contributes to poor morale and self-efficacy issues (Abraham et al., 2004; Bass, 2002). This role of the nurse as being critical in patient throughput needs to be adequately articulated to the population, management, and associated physicians (Abraham et al., 2004). With the dire status of the industry and the grave projections of the impact on patient care, more and more health-care organization administrators recognize the need to establish a vision of change for nurses, articulate how they fit into the care process, and explain why it is important for nurses to accept the technologies that enable efficiencies. Bringing physicians into the fold to internalize and articulate this message will also undoubtedly promote acceptance of the technology among the nursing population since they have a strong need for external affirmation. The vision needs to manifest into useful but nonobtrusive IT and IS training for nurses in order to improve computer self-efficacy and increase the likelihood that the technology will be used as intended to produce desired efficiencies. Nurses commented that the majority of the physicians they work with do not appreciate their services, which contribute to the nurses rejecting many technologies that management presents to them because they feel that the technologies are being thrust upon them in an effort to make the physicians' jobs less demanding and not for improving nursing work conditions (Abraham et al., 2004). The technology can be the main facilitator to empower nurses as the principal information managers in patient care. One area of future research is to explore more thoroughly how technology is facilitating a change in the authoritative structure between physicians and nurses. The hope is for physicians to see nurses as more competent and capable because of the ability to more readily address questions of them made possible with the use of IT and IS (Abraham et al., 2004).

FUTURE TRENDS

A future trend is the exploration of nursing decision support systems comparable to those developed for physicians, but with a focus on nursing-related patient care tasks that have elements of diagnoses. Another area for future research is the incorporation of project management skills and facilitating IS and IT into nursing curricula in order to aid nurse managers and clinical nurses with a means for balancing patient loads with fewer nurse resources.

CONCLUSION

The primary contribution of this chapter is to promote visibility of the needs of the nursing community regarding electronic information exchange, manipulation, and management that can be addressed in healthcare information systems research. Typically, IS research focus groups are geared toward physicians; however, nurses are responsible for the bulk of documentation for each medical intervention in each step of the patient care system. Physicians rely heavily on the information supplied by nurses concerning the historical documentation of charting patients' ailments, effectiveness of prescribed regimens, and anomalies in their condition (Abraham et al., 2004). Focus has been placed on using technology to streamline processes and automate as much as possible to avoid human documentation error. However, healthcare IS research has frequently underexplored nurses or not recognized them as the principal personnel who supply the information to the system, whether the system is manual or automated. IS researchers can help to uncover nuances in the nature of nursing and design IT and IS to better support the way nurses work in hopes of helping our society achieve its goal of improving the quality of patient care.

REFERENCES

Abbott, P. (2006). Innovations beyond the lab: Putting information into the hands of those who care. In *Proceedings of the 2006 American Nursing Informatics Association Conference*, Nashville, Tennessee.

Abraham, C., Watson, R.T., & Boudreau, M.C. (forthcoming). Ubiquitous access: Patient care on the Frontlines. *Communications of the ACM*.

Abraham, C, Watson R.T., Boudreau, M.C., & Goodhue, D. (2004) *Patient care and safety at the frontlines: Nurses' experiences with wireless computing*. Publication Series for IBM Center for Healthcare Management

American Nurses Informatics Association. (2006). *Description for Bayada Award for technological innovation in nursing education and practice*. Retrieved from http://www.ania.org/Bayada.htm

Bass, J. (2002). *It's not just nurses anymore: Survey of three other health occupations finds patient safety at risk due to understaffing*. Retrieved from http://www.aft.org/press/2002/041102.html

Bostrom, R., & Heinen, J. (1977). MIS problems and failures: A socio-technical perspective. Part I: The causes. *MIS Quarterly, 1*(3), 17–32.

Bove, L. (2006). Implementing a system with clinical transformation in mind. In *Proceedings of the 2006 American Nursing Informatics Association Conference,* Nashville, Tennessee.

Buchan, J., & Calman, L. (2004). *The global shortage of registered nurses: An overview of issues and actions.* International Council of Nurses. Retrieved from http://www.icn.ch/global/shortage.pdf

Council on Systemic Interoperability. (2006). *Interviews from practitioners speaking about ending the document game in U.S. healthcare*. Retrieved from http://endingthedocumentgame.gov/video.html

Institute of Medicine (IOM). (2000). To err is human: Building a safer health system. The National Academic Press. Retrieved from http://www.nap.edu/books/0309068371/html/

International Council of Nurses. (2006). *International classification for nursing practice (ICNP) fact sheet*. Retrieved from http://www.icn.ch/icnp.htm

Joint Commission on Accreditation of Healthcare Organizations (JCAHO). (2004). *Healthcare at the crossroads: Strategies for addressing the evolving nursing crisis.* Retrieved from http://www.jointcommission.org/NR/rdonlyres/5C138711-ED76-4D6F-909F-B06E0309F36D/0/health_care_at_the_crossroads.pdf

Patistea, E. (1999). Nurses' perceptions of caring as documented in theory and research. *Journal of Clinical Nursing, 8*(5), 487–495.

Reinertsen, J., Pugh, M., & Bisognano, M. (2005). *Seven leadership leverage points for organization-level improvement in healthcare.* Retrieved from http://www.delmarvafoundation.org/html/culture_club/docs/Seven%20Leadership%20Leverage%20Points.pdf

Romano, C. (2006). Improving the health of the nation: The promise of health information technology. In *Proceedings of the 2006 American Nursing Informatics Association Conference*, Nashville, Tennessee.

Sand, A. (2003). Nurses' personalities, nurses-related qualities and work satisfaction: A ten year perspective. *Journal of Clinical Nursing, 12*, 177–187.

TelecomWeb. (2005). *U.S. healthcare providers seen growing IT budgets by over 10 percent*. Retrieved from http://www.telecomweb.com/news/1086273449.htm

Wiley-Patton, S., & Malloy, A. (2004). Understanding healthcare professionals' adoption and use of IT. In *Proceedings of the Tenth Americas Conference on Information Systems*, New York.

KEY TERMS

Charting: The process of monitoring the patient and recording vital statistics, effectiveness of medical intervention, and overall patient wellness, an ongoing process across nursing shifts.

Electronic Medical or Health Records: Electronic version of the manual chart that outlines historical medical information pertaining to one individual.

Medication Administration: The process of validating medication regimens with prescriptions, delivering the medication, and documenting the effectiveness.

Mobile or Wireless Computing: The use of mobile, untethered devices to access information stored in information systems or communicate via wireless networks.

Nursing: Encompasses autonomous and collaborative care of individuals of all ages, families, groups, and communities, sick or well and in all settings. Nursing includes the promotion of health, prevention of illness, and the care of ill, disabled, and dying people. Advocacy, promotion of a safe environment, research, participation in shaping health policies and patient and health systems management, and education are also key nursing roles.

Systemic Interoperability: Comprehensive network of privacy-protected systems of electronic personal health information developed around the healthcare consumer to facilitate wellness and the safe delivery of optimal healthcare.

Throughput Flow: The process of managing when, where, and how patients are moved about in the patient care system across units within the hospital for labs, procedures, and so forth.

Triage: The process of taking vital statistics, documenting impending problems, and making initial assessments to categorize criticality of the emergent condition.

Regularity Analysis of the Magnetoencephalogram Background Activity in Alzheimer's Disease Patients Using Auto Mutual Information

Carlos Gómez
University of Valladolid, Spain

Roberto Hornero
University of Valladolid, Spain

Daniel Abásolo
University of Valladolid, Spain

Alberto Fernández
Complutense University of Madrid, Spain

Javier Escudero
University of Valladolid, Spain

INTRODUCTION

Alzheimer's disease (AD) is the most common form of dementia (Bird, 2001), a group of conditions that gradually destroys brain cells and leads to progressive decline in mental function. This irreversible brain disorder is characterized by neuronal loss and the appearance of neuritic plaques containing amyloid-β-peptide and neurofibrillary tangles (Cummings, Pike, Shankle, & Cotman, 1996). Approximately 50–60% of patients with dementia over 65 years are clinically related to AD and the number of patients is expected to increase continuously (Lahiri, Farlow, Greig, & Sambamurti, 2002). A differential diagnosis with other types of dementia and with major depression is used. It can include Mini-Mental Statues Examination (MMSE) (Folstein, Folstein, & McHugh, 1975), magnetic resonance imaging, computerized axial tomography, positron emission tomography, and verbal tests. Nevertheless, a definite diagnosis of AD is only possible by necropsy. In order to complete the diagnosis, nonlinear analysis of the brain recordings might be used.

Magnetoencephalograpy (MEG) is a noninvasive technique that allows recording the magnetic fields produced by brain activity. SQUID (Superconducting QUantum Interference Device) sensors immersed in liquid helium at 4.2 K are used to detect the extremely weak brain magnetic signals. MEG provides an excellent temporal resolution, orders of magnitude better than in other methods for measuring cerebral activity, as magnetic resonance imaging, single-photon-emission computed tomography and positron-emission tomography (Hämäläinen, Hari, Ilmoniemi, Knuutila, & Lounasmaa, 1993). A good spatial resolution is also provided, although this depends on the source configuration (Hari & Forss, 1999). Moreover, this technique is independent of any reference point. In addition, magnetic fields are not distorted by the resistive properties of the skull (Hämäläinen et al., 1993). Nevertheless, the recordings are very sensitive to external artifacts. Thus, the signals must be acquired in a magnetically shielded room.

In this preliminary study, we examined the MEG background activity in patients with probable AD, and in age-matched control subjects using auto mutual information (AMI). Our purpose is to test the hypothesis that an abnormal type of nonlinear dynamics is associated with AD.

BACKGROUND

In the last years, brain activity of AD patients has been analyzed with several nonlinear methods. The most widely used are the correlation dimension (D_2) and the first Lyapunov exponent (L_1). D_2 is a measure of the system dimensional complexity (Grassberger & Procaccia, 1893a, 1893b) while L_1 is a dynamic complexity measure that describes the divergence of trajectories starting at nearby initial states (Jeong, 2004). Jeong, Kim, and Han (1998) showed that AD patients' EEGs exhibit significantly lower D_2 and L_1 than the EEGs of control subjects in many channels. Van Cappellen van Walsum, Pijnenburg, Berendse, van Dijk, Know, Scheltens, and Stam (2003) used MEG data and computed the D_2 in different frequency bands. In the broad band 0.5–40 Hz, the mean D_2 was lower in AD patients compared with control subjects. Nevertheless, there are several drawbacks in using these nonlinear methods. Reliable estimations of D_2 and L_1 require a large quantity of data (Eckmann & Ruelle, 1992), and stationary and noise-free time series (Grassberger & Procaccia, 1983a). These assumptions cannot be achieved for physiological data. Moreover, the computational cost is high, especially for great amount of data.

Therefore, other nonlinear methods which avoid these disadvantages are necessary to study the MEG background activity. Lempel-Ziv complexity measures the complexity of finite sequences, and is related to the number of distinct substrings and the rate of their occurrence along the sequence (Lempel & Ziv, 1976). Recent studies have applied this nonlinear method to EEG (Abásolo, Hornero, Gómez, García, & López, 2006b) and MEG signals (Gómez, Hornero, Abásolo, Fernández, & López, 2006) of AD patients and elderly control subjects, finding significant differences between both groups in certain regions of the brain. Other methods measure the irregularity of a signal, as the approximate entropy and the sample entropy. With both methods, an increased regularity was found in the EEG of AD patients (Abásolo, Hornero, Espino, Álvarez, & Poza, 2006a; Abásolo, Hornero, Espino, Poza, Sánchez, & de la Rosa, 2005). Finally, the interdependencies between MEG signals in six frequency bands were studied with synchronization likelihood (Stam, Jones, Manshanden, van Cappellen van Walsum, Montez, Verbunt, de Munck, van Dijk, Berendse, & Scheltens, 2006).

In this study, another nonlinear method, the auto mutual information (AMI), has been used to analyze the MEG signals. Mutual information (MI) has been used to study different types of diseases and brain states. It has been shown that MI might be useful to predict the response to anesthesia (Huang, Yu, Ju, & Cheng, 2003). Xu, Liu, Liu, and Yang (1997) computed the complexity of cross-mutual information (CMI) functions among eight EEG channels for four different functional states: awake with opened and closed eyes, light sleep, and deep sleep. Na, Jin, Kim, and Ham (2002) estimated the decreased rate of AMI and the CMI values from the EEG of ten schizophrenic patients and ten normal controls. Both nonlinear methods were also applied to study the EEG activity in AD (Jeong, Gore, & Peterson, 2001a). They found that AMI profiles decreased more slowly with time delay throughout the brain in AD patients than in control subjects. The CMI analyses showed a significantly decreased transmission of information between pairs of AD patients' electrodes.

MATERIALS AND METHODS

Subjects and MEG Recording

MEG signals were recorded with a 148-channel whole head magnetometer (MAGNES 2500 WH, 4D Neuroimaging) in a magnetically shielded room. The subjects lay on a patient bed, in a relaxed state, and with their eyes closed. Five minutes of recording were acquired at a sampling frequency of 678.17 Hz. Then, these recordings were down-sampled to 169.549 Hz. Artifact-free epochs of 20 seconds (3392 data points) were selected and filtered with a band-pass filter (0.5-40 Hz).

The MEG data were acquired from 30 subjects. 15 patients (three men and 12 women) fulfilling the criteria of probable AD (age = 71.13 ± 8.52 years, mean ± standard deviation SD) participated in the present study. All of them were recruited from the Asociación de Familiares de Enfermos de Alzheimer (AFAL). The patients were diagnosed according to the criteria of the National Institute of Neurological and Communicative Disorders and Stroke and the AD and Related Disorders Association (NINCDS-ADRDA) (McKhann, Drachman, & Folstein, 1984). Functional assessment staging procedure (FAST) was used for assessment of functional ability and MMSE for measuring the cognitive status of the subjects. The MMSE and FAST scores were 18.07 ± 4.18 and 4.00 ± 0.38 (mean ± SD), respectively. None of the patients used any kind of medication that could have an influence on the MEG.

MEGs were also obtained from 15 age-matched control subjects (seven men and eight women, age = 70.73 ± 8.29 years, MMSE = 29.13 ± 0.99, FAST = 1.67 ± 0.49, mean ± SD). Sex, age, MMSE, and FAST scores of all subjects are shown in Table 1. The local ethics committee approved the study. All control subjects and all caregivers of the demented patients gave their informed consent for the participation in the current research.

Auto Mutual Information

Mutual information (MI) provides a measure of both the linear and nonlinear statistical dependencies between two time series (Jeong et al., 2001a). It can be defined

Table 1. Sociodemographic data of AD patients and control subjects

Id.	Age	Sex	MMSE	FAST	Id.	Age	Sex	MMSE	FAST
Alz-1	71	Female	15	4	Con-1	84	Male	29	2
Alz-2	83	Male	10	5	Con-2	61	Female	29	2
Alz-3	56	Female	14	4	Con-3	70	Female	30	2
Alz-4	64	Female	15	4	Con-4	64	Female	30	1
Alz-5	59	Female	20	4	Con-5	60	Male	30	1
Alz-6	60	Male	16	4	Con-6	63	Female	30	1
Alz-7	72	Female	15	4	Con-7	73	Male	29	1
Alz-8	71	Female	15	4	Con-8	69	Female	29	1
Alz-9	75	Female	22	4	Con-9	56	Female	27	2
Alz-10	82	Female	21	4	Con-10	79	Male	29	2
Alz-11	72	Female	17	4	Con-11	79	Male	30	2
Alz-12	80	Male	24	4	Con-12	75	Female	29	2
Alz-13	77	Female	21	4	Con-13	74	Male	30	2
Alz-14	73	Female	23	4	Con-14	78	Male	27	2
Alz-15	72	Female	24	3	Con-15	76	Female	29	2
Mean ± SD	71.13 ± 8.52	–	18.07 ± 4.18	4.00 ± 0.38	Mean ± SD	70.73 ± 8.29	–	29.13 ± 0.99	1.67 ± 0.49

as the amount of information gained about one signal from the measurement of another (Xu et al., 1997). Furthermore, MI between two time series is zero when those series are completely independent, while MI has a maximum value if both series are equal.

We can compute the time-delayed mutual information between X and Y as follows (Jeong et al., 2001a):

$$I_{XY_\tau} = \sum_{x(t),y(t+\tau)} P_{XY_\tau}\left[x(t), y(t+\tau)\right] \cdot \log_2\left\{\frac{P_{XY_\tau}\left[x(t), y(t+\tau)\right]}{P_X\left[x(t)\right]\cdot P_{Y_\tau}\left[y(t+\tau)\right]}\right\} \quad (1)$$

where $P_X\left[x(t)\right]$ and $P_{Y_\tau}\left[y(t+\tau)\right]$ are the normalized histograms of the distribution of values observed for $x(t)$ and $y(t+\tau)$, while $P_{XY_\tau}\left[x(t), y(t+\tau)\right]$ is the joint probability density for the measurements of $x(t)$ and $y(t+\tau)$. This is the CMI and quantifies the information transmitted from one signal to another (Jeong et al., 2001a). Applied to MEG signals, the CMI measures the quantity of information transmitted between certain areas of the brain.

To calculate the AMI, the MI between $x(t)$ and $x(t+\tau)$, Equation 1 can be rewritten as:

$$I_{XX_\tau} = \sum_{x(t),x(t+\tau)} P_{XX_\tau}\left[x(t), x(t+\tau)\right] \cdot \log_2\left\{\frac{P_{XX_\tau}\left[x(t), x(t+\tau)\right]}{P_X\left[x(t)\right]\cdot P_{X_\tau}\left[x(t+\tau)\right]}\right\} \quad (2)$$

The AMI quantifies the mean predictability of $x(t+\tau)$ from $x(t)$ (Jeong et al., 2001a). To calculate AMI from experimental data is necessary to estimate the probability distributions of X and X_τ and the joint probability density P_{XX_τ}. These probabilities can be estimated constructing the histograms of the time series. For a given epoch length, the use of larger sampling bins produces more accurate estimates of the average probability, but the estimate of the joint probability distribution is too flat, underestimating the AMI. On the other hand, the use of smaller bins is better to indicate changes in the joint probability over short distances, but it produces fluctuation because the sample size is small, overestimating the AMI (Jeong et al., 2001a). In previous studies, 64 bins has been used to construct the histograms because this value provides stable estimations (Jeong et al., 2001a; Min, Jin, Kang, Lee, Kang, Lee, & Sakamoto, 2003; Na et al., 2002). In our research work, we have estimated the joint probability density using histograms of eight, 16, 32, and 64 bins.

We evaluated the AMI of all channels over a time delay from 0 to 0.5 seconds. In order to normalize the AMI profiles, they have been divided by the AMI value at $\tau = 0$. Hence, AMI values at a zero time delay are always one. The slope of the AMI profile was estimated by a line that fits the data in a least-squares sense. This slope was calculated from $\tau = 0$ to the first minimum value of the profile. Palus (1993) suggested that the decline rate of the AMI was positively correlated with the entropy. As entropy measures quantify the regularity (Costa, Goldberger, & Peng, 2002; Pincus, Gladstone, & Ehrenkranz, 1991), this decrease rate can be used to estimate the regularity of a signal.

Finally, a statistical analysis was carried out separately for each channel. Student's *t*-test was used to evaluate the statistical differences between the slopes of the AMI for AD patients and control subjects.

RESULTS

AMI was estimated for the 148 MEG channels, with a maximum time delay of 0.5 seconds, and eight, 16, 32, and 64 bins for the construction of the histograms. Figure 1 illustrates the normalized AMI profiles for the control subjects and the AD patients at squid A1. This figure shows that the AMI values decrease quickly for low τ values, and then become stable when the values of the time delay τ increase. As can be noticed, the AMI profile of the AD group declines more slowly than the profile of the control group. This behavior is the same for all channels, indicating that AD patients' MEGs are more regular than control subjects' recordings. Our results suggest that there are differences between the nonlinear dynamics of AD patients and control subjects' MEGs.

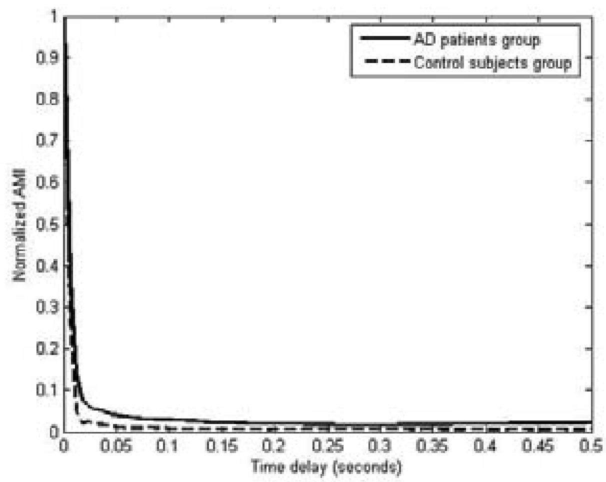
Moreover, these differences are statistically significant ($p < 0.01$, Student's *t*-test) in 133 channels for histograms of eight bins, in 132 channels for 16 bins, in 130 for 32 bins, and in 127 channels if the histograms are constructed with 64 bins. Figure 2 illustrates a squid diagram with the results of the Student's *t*-test using eight bins for the construction of the histograms. Although previous articles suggested the use of 64 bins to construct the histograms (Jeong et al., 2001a; Min et al., 2003; Na et al., 2002), our study shows that with a lower number of bins, the results get better.

DISCUSSION

This study presents the AMI as a method to analyze the MEG background activity in 15 patients with probable AD, and in 15 control subjects. AMI estimates the degree to which $x(t+\tau)$ can be predicted from $x(t)$. Moreover, the rate of decrease of the AMI with increasing τ is a measure of the MEGs regularity. For our pilot study, we used values of $\tau_{max} = 0.5$ seconds, and eight, 16, 32, and 64 bins to construct the histograms.

Traditional nonlinear techniques, as L_1 or D_2, have been widely applied to study the brain background activity (Jeong et al., 1998; van Cappellen van Walsum

Figure 1. Normalized AMI curves of the 15 control subjects and the 15 patients with probable AD at channel A1 (64 bins)



et al., 2003). Nevertheless, for a suitable estimation of these measures, stationary and noise-free signals, and a large number of data points, are needed (Eckmann & Ruelle, 1992; Grassberger & Procaccia, 1983a). These assumptions cannot be achieved for physiological data, as EEG and MEG. On the other hand, AMI does not require a large number of data points to be reliably estimated, and can be applied to nonstationary time series (Jeong et al., 2001a). Thus, this measure is much better suited for MEG complexity analysis than traditional nonlinear techniques.

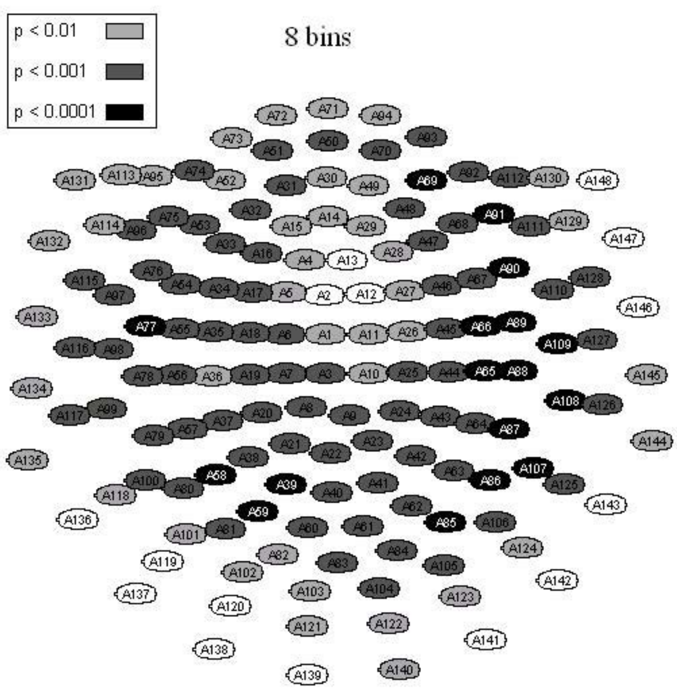
We found that the absolute values of the average AMI decrease rate are lower in AD patients in all channels. The differences were statistically significant ($p < 0.01$) in 133, 132, 130, and 127 channels for histograms of eight, 16, 32, and 64 bins, respectively. Because AD patients' group and control subjects' group were carefully matched for age, the significantly reduced AMI decline rate may well represent the cognitive dysfunction in AD.

Our results agree with previous studies of the EEG/MEG background activity in AD patients (Abásolo et al., 2005; Abásolo et al., 2006a; Jeong et al., 2001a). These studies suggested a loss of irregularity of the brain activity in the patients with probable AD. Jeong et al. (2001a) found that AMI profiles decrease with time delay slower in AD patients' EEGs than in control subjects' ones. Other study showed that sample entropy values were significantly lower in the EEG of AD patients at parietal and occipital regions (Abásolo et al., 2006a). With another nonlinear method, approximate entropy, significant differences were found at P3 and P4 EEG electrodes (Abásolo et al., 2005). The implications of this decreased irregularity in AD patients are not clear. It might arise from neuronal death, deficiency of neurotransmitters like acetylcholine, and/or loss of connectivity of local neuronal networks (Jeong, 2004).

FUTURE TRENDS

We found that an abnormal type of nonlinear dynamics is associated with AD. However, changes in EEG/MEG activity also appear in other pathological states as schizophrenia (Na et al., 2002), vascular dementia (Jeong, Chae, Kim, & Han, 2001b) and epilepsy (Jing & Takigawa, 2001). Therefore, we need a larger database, including recordings of these pathological states to confirm the performance of our method.

Figure 2. Squid diagrams with the results of the Student's t-test using eight bins for the construction of the histograms



The results of our study show that there are important statistical significant differences between controls and AD patients detectable using the methodology presented in our article. Nevertheless, our results do not show if AMI can detect a gradation of the disease process. Therefore, future studies are necessary to analyse MEGs from patients with different stages of AD and with mild cognitive impairment.

CONCLUSION

Our results show that irregularity, measured via AMI, is lower in AD patients' MEGs than in control subjects' ones. Moreover, the differences were statistically significant for most channels ($p < 0.01$, Student's *t*-test). This study suggests that cognitive dysfunction in AD is associated with an increased regularity in the MEG recordings.

In addition to this, we have shown that the AMI decrease rate could be a good method to differentiate between AD patients and elderly control subjects. Nevertheless, this is a preliminary study and a larger database is needed to confirm our results.

REFERENCES

Abásolo, D., Hornero, R., Espino, P., Álvarez, D., & Poza, J. (2006a). Entropy analysis of the EEG background activity in Alzheimer's disease patients. *Physiological Measurement, 27*, 241–253.

Abásolo, D., Hornero, R., Espino, P., Poza, J., Sánchez, C. I., & de la Rosa, R. (2005). Analysis of regularity in the EEG background activity of Alzheimer's disease patients with approximate entropy. *Clinical Neurophysiology, 116*, 1826–1834.

Abásolo, D., Hornero, R., Gómez, C., García, M., & López, M. (2006b). Analysis of EEG background activity in Alzheimer's disease patients with Lempel–Ziv complexity and Central Tendency Measure. *Medical Engineering & Physics, 28*, 315–322.

Bird, T. D. (2001). Harrison's principles of internal medicine. In E. Braunwald, A. S. Fauci, D. L. Kasper, S. L. Hauser, D. L. Longo, & J. L. Jameson, (Eds.), *Alzheimer's Disease and Other Primary Dementias* (pp. 2391–2399). New York: The McGraw-Hill Companies Inc.

Costa, M., Goldberger, A. L., & Peng, C. K. (2002). Multiscale entropy analysis of complex physiologic time series. *Physical Review Letters, 89*, 068102.

Cummings, B. J., Pike, C. J., Shankle, R., & Cotman, C. W. (1996). Betaamyloid deposition and other measures of neuropathology predict cognitive status in Alzheimer's disease. *Neurobiology of Aging, 17*, 921–933.

Eckmann, J. P., & Ruelle, D. (1992). Fundamental limitations for estimating dimensions and Lyapunov exponents in dynamical systems. *Physica D, 56*, 185–187.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). Mini-mental state: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12*, 189–198.

Gómez, C., Hornero, R., Abásolo, D., Fernández, A., & López, M. (2006). Complexity analysis of the magnetoencephalogram background activity in Alzheimer's disease patients. *Medical Engineering & Physics, 28*, 851–859.

Grassberger, P., & Procaccia, I. (1983a). Characterization of strange attractors. *Physical Review Letters, 50*, 346–349.

Grassberger, P., & Procaccia, I. (1983b). Estimation of the Kolmogorov entropy from a chaotic signal. *Physical Review A, 28*, 2591–2593.

Hämäläinen, M., Hari, R., Ilmoniemi, R. J., Knuutila, J., & Lounasmaa, O. V. (1993). Magnetoencephalograpy – theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of Modern Physics, 65*, 413–497.

Hari, R., & Forss, N. (1999). Magnetoencephalograpy in the study of human somatosensory cortical processing. *Philosophical Transactions of the Royal Society B, 354*, 1145–1154.

Huang, L., Yu, P., Ju, F., & Cheng, J. (2003). Prediction of response to incision using the mutual information of electroencephalogram during anaesthesia. *Medical Engineering & Physics, 25*, 321–327.

Jeong, J. (2004). EEG dynamics in patients with Alzheimer's disease. *Clinical Neurophysiology, 115*, 1490–1505.

Jeong, J., Chae, J. H., Kim, S. Y., & Han, S. H. (2001b). Nonlinear dynamic analysis of the EEG in patients with Alzheimer's disease and vascular dementia. *Journal of Clinical Neurophysiology, 18*, 58–67.

Jeong, J., Gore, J. C., & Peterson, B. S. (2001a). Mutual information analysis of the EEG in patients with Alzheimer's disease. *Clinical Neurophysiology, 112*, 827–835.

Jeong, J., Kim, S. Y., & Han, S.-H. (1998). Nonlinear dynamical analysis of the EEG in Alzheimer's disease with optimal embedding dimension. *Electroencephalography and Clinical Neurophysiology, 106*, 220–228.

Jing, H., & Takigawa, M. (2001). Comparison of human ictal, interictal and normal nonlinear component analyses. *Clinical Neurophysiology, 111*, 1282–1292.

Lahiri, D. K., Farlow, M. R., Greig, N. H., & Sambamurti, K. (2002). Current drugs targets for Alzheimer's disease treatment. *Drug Development Research, 56*, 267–281.

Lempel, A., & Ziv, J. (1976). On the complexity of finite sequences. *IEEE Transactions on Information Theory, 22*, 75–81.

McKhann, G., Drachman, D., & Folstein, M. (1984). Clinical diagnosis of Alzheimer's disease: Report of NINCDS-ADRDA work group under the auspices of department of health and human services task force on Alzheimer's disease. *Neurology, 34*, 939–944.

Min, B. C., Jin, S. H., Kang, I. H., Lee, D. H., Kang, J. K., Lee, S. T., & Sakamoto, K. (2003). Analysis of mutual information content for EEG responses to odor stimulation for subjects classified by occupation. *Chemical Senses, 28*, 741–749.

Na, S. H., Jin, S. -H., Kim, S. Y., & Ham, B. -J. (2002). EEG in schizophrenic patients: Mutual information analysis. *Clinical Neurophysiology, 113*, 1954–1960.

Palus, M. (1993). Coarse-grained entropy rates for characterization of complex time series. *Physica D, 93*, 64–77.

Pincus, S. M., Gladstone, I. M., & Ehrenkranz, R. A. (1991). A regularity statistic for medical data analysis. *Journal of Clinical Monitoring, 7*, 335–345.

Stam, C. J., Jones, B. F., Manshanden, I., van Cappellen van Walsum, A. M., Montez, T., Verbunt, J. P. A., de Munck, J. C., van Dijk, B. W., Berendse, H. W., & Scheltens, P. (2006). Magnetoencephalographic evaluation of resting-state functional connectivity in Alzheimer's disease. *Neuroimage, 32*, 1335–1344.

van Cappellen van Walsum, A. M., Pijnenburg, Y. A. L., Berendse, H. W., van Dijk, B. W., Know, D. L., Scheltens, P., & Stam, C. J. (2003). A neural complexity measure applied to MEG data in Alzheimer's disease. *Clinical Neurophysiology, 114*, 1034–1040.

Xu, J., Liu, Z. -R., Liu, R., & Yang, Q. -F. (1997). Information transmission in human cerebral cortex. *Physica D, 106*, 363–374.

KEY TERMS

Alzheimer's Disease (AD): Progressive, degenerative, and irreversible brain disorder characterized by the appearance of neuritic plaques and neurofibrillary tangles.

Auto Mutual Information (AMI): Mutual information between a signal and its time-delayed version.

Electroencephalogram (EEG): Neurophysiologic recording of the brain electrical activity by electrodes placed on the scalp.

Magnetoencephalogram (MEG): Neurophysiologic recording of the magnetic fields produced by brain activity using extremely sensitive devices called SQUIDs.

Mutual Information (MI): Amount of information gained about one signal from the measurement of another.

Nonlinear Analysis: Broad, interdisciplinary field characterized by a mixture of analysis, topology, and geometry.

Signal Processing: Extraction of information from complex signals, generally by conversion of the signals into digital form followed by analysis using different algorithms.

A Rehabilitative Eye Tracking-Based Brain-Computer Interface for the Completely Locked-In Patient

Ziad O. Abu-Faraj
American University of Science and Technology, Lebanon

Jean-Louis D. Heneine
American University of Science and Technology, Lebanon

Habib C. Bou Sleiman
American University of Science and Technology, Lebanon

Maya J. Mashaalany
American University of Science and Technology, Lebanon

Waleed M. Al Katergi
American University of Science and Technology, Lebanon

INTRODUCTION

Despite contemporary advancements in healthcare, conventional methods of treatment have more than once proved ineffective in curing certain classes of neurological disorders, especially those where the central nervous system has been subjected to severe irreversible damage. Such disorders include, but are not limited to, cerebral palsy, cerebral aneurysm, traumatic brain injury, stroke, apraxia, and aphasia. In certain severe cases, more than one disorder might affect the same patient, making him or her completely locked-in—that is, cognitively intact, but unable to move or communicate (Kennedy & Adams, 2003).

A completely locked-in patient usually suffers from symptoms that include complete inability of controlling any voluntary muscles in the body, apart from those needed for eye movements and blinking. Nonetheless, the patient is capable of reasoning, thinking, as well as preserving all signs of consciousness. Moreover, normal sleep and wake cycles persist throughout the locked-in state. In addition, as is the case of most paralyzed people, a locked-in patient has the tendency to develop muscle atrophy and spasticity. Figure 1 illustrates a sketch of a locked-in patient.

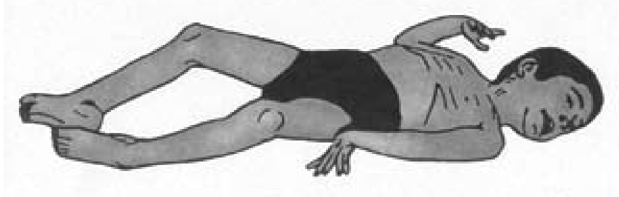
There are vast regions around the world where the completely locked-in patient is doomed to social rejection—in particular, when conventional rehabilitation methods fail to lessen the severity of the patient's handicap. For example, in the absence of motor control, it is practically impossible to use sign language or common input devices to interface with a computer system in order to communicate through spelling or expression-building software (Betke, Gips, & Fleming, 2002).

This article describes the design and development of a low-cost eye tracking-based brain-computer interface system for the rehabilitation of the completely locked-in patient having an intact ocular motor control to serve as an alternative means of communication (Abu-Faraj, Mashaalany, Bou Sleiman, Heneine, & Katergi, 2006). The developed system has been designed according to the following criteria: low cost, low processing power, simplicity of operation, little training requirements, minimal disturbance to the patient, and ease of customization to any mother tongue.

BACKGROUND

Brain-computer interface (BCI) methods for the rehabilitation of patients with neural disorders, similar to those previously listed, have been widely published. BCI systems are primarily classified according to the input technique utilized. Commonly reported techniques include electromyography (EMG) (Nagata, Yamada, & Magatani, 2004), electroencephalography (EEG) (Pfurtscheller, Müller-Putz, Pfurtscheller, & Rupp, 2005), electrooculography (EOG) (Barea, Boquete,

Figure 1. A sketch of a locked-in child



Mazo, & López, 2002), eye tracking (Amarnag, Kumaran, & Gowdy, 2003; Betke, Gips, & Fleming, 2002; Böhme & Barth, 2005; Cui, Ma, Wang, Tan, & Sun, 2004; Lee & Galiana, 2005; Li, Winfield, & Parkhurst, 2005; Yu & Eizenman, 2004), and other custom-designed input devices (Chen, 2001; Chen, Tang, Chang, Wong, Shih, & Kuo, 1999). The decision about which input technique to be used depends on the specific case being studied (Kennedy & Adams, 2003).

The EMG-based BCI requires that the patient be able to control at least one to three distinct muscles in order to be functionally effective, while the head-mounted tilt sensor method (Chen, 2001) can be used only if the patient has the voluntary ability of rotating his or her head. Accordingly, these two methods are precluded from consideration, since a locked-in patient does not possess any perceptible movements or EMG activity. Consequently, BCI methods employing eye tracking, EEG, or EOG remain possible for applications involving a locked-in patient (Kennedy & Adams, 2003). Nevertheless, EEG systems are expensive, sophisticated, cumbersome, and necessitate extensive training, while EOG methods suffer from inaccuracies at the extreme position of the pupil—particularly, small angle displacements (less than 2˚) are difficult to record, whereas large eye movements (greater than 30˚) do not produce bioelectric amplitudes that are strictly proportional to eye position (Clark, 1998). Hence, eye tracking systems would be more adequate for such applications, and are easier to setup.

THE EYE TRACKING BCI SYSTEM

The core of this work was to implement a fully functional rehabilitative system, while maintaining the development cost to a bare minimum. For this reason, the design specifications and requirements of the hardware were not only reduced, but also compensated for by the software program.

Hardware Description

A Genius® (KYE Systems Corp., Chung, Taipei Hsien, Taiwan, R.O.C.) Web cam was utilized and found suitable in terms of resolution, light sensitivity, and acceptable cost. The external packaging of the camera was redesigned, thereby substantially reducing its weight and volume, in order for it to be mounted on the patient's head. This alteration produced a headset consisting of a simple mechanical framework that separates the camera lens from the patient's eye by about five centimeters. The Web cam was connected to a notebook computer via a regular USB 2.0 port. The computer used in the development and testing of this system was a Toshiba Satellite™ M70-122 (Toshiba America, Inc., New York, NY, USA) notebook with a 1.73 GHz Centrino™ (Intel, Santa Clara, CA, USA) processor, 512 MB of RAM, and a 15 in. widescreen. Figure 2 shows a normal child instrumented with the eye tracking system.

Figure 2. A normal child instrumented with the eye tracking system

Software Description

A Matlab® (The MathWorks, Inc., Natick, MA, USA) program was developed to analyze the Web cam images of the contour of the eye in real-time. This program was implemented under Matlab® version 6.5, release 13, and consists of a series of image processing routines that are designed to locate the position of the center of the pupil. The point of gaze is localized as a pair of rectangular coordinates (x and y). The pseudocode lines presented in Figure 3 describe the algorithm that was exclusively developed to be used in this application. This algorithm proved to be efficient in terms of accuracy and processing speed. Furthermore, it was capable of processing around four frames-per-second on the previously designated computer system, a frame rate that is adequate for this specific application.

Upon acquiring the image of the eye (A), the program first converts it to an eight-bit grayscale intensity image, and then sweeps a black search window (w x h) over the entire surface of the frame (W x H). At each position (i, j) of the window, the program computes the error between the window and the corresponding area of the frame. The area that produces the lowest error is hence considered to be the darkest region of the image. Subsequently, this area is the approximated position of the pupil, which is marked at the center of the search window. This process is illustrated in Figure 4. Prior to running this pupil-allocating algorithm, the program cancels the effect of any white spot caused by reflection of light on the eye by eliminating high-intensity pixels. At the end of this algorithm, the coordinates of the pupil are passed to the next stage of processing, where interpretation of the data takes place.

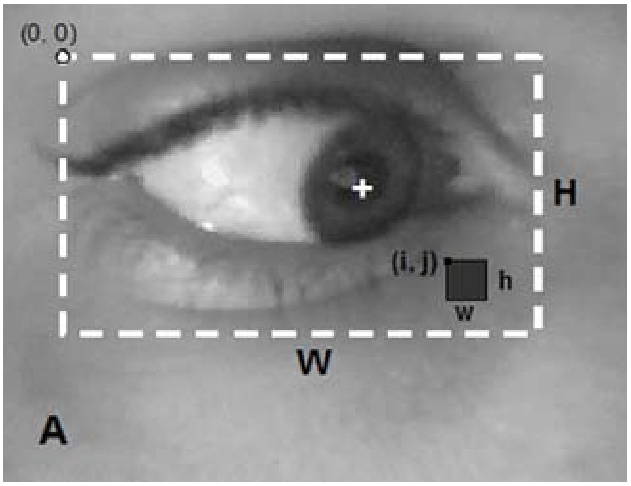
The next stage of the software system is a graphical user interface (GUI) implemented under the JAVA™ platform (Sun Microsystems, Inc., Santa Clara, CA, USA). This GUI consists of various searchable pages, each of which contains a (2 x 3) grid of large buttons showing a collection of pictograms. Pictograms are simple, understandable images that represent basic human daily needs. As the locked-in patient focuses on a particular pictogram, the program determines the point of gaze and measures the duration that the patient spends looking at that pictogram. Three seconds were considered ample to ascertain the patient's intention. However, this duration can be adjusted by the computer operator.

Upon the confirmation of the patient's selection of a pictogram, the program obtains the corresponding wave file from the computer hard disk and plays it back to the patient's attendant through the built-in standard wave output device. The wave files can be recorded independently of the system, and in any desired language. To ensure that no erroneous executions are made, all buttons are initially *disabled,* except for a particular

Figure 3. A pseudocode of the algorithm used for the detection of the patient's pupil

```
proc locate (A: array [1..H][1..W] of double)
    for i ← 1 to (H – h) do
        for j ← 1 to (W – w) do
            for m ← i to (i + h) do
                for n ← 1 to (j + w) do
                    error ← error + A [m, n];
            if (error <= minError)
                minError ← error;
                x ← (i + h/2);
                y ← (j + h/2) ;
            error ← 0;
    return x, y;
```

Figure 4. A visual illustration of the mechanism used by the algorithm to determine the approximated pupil position; A = image of the eye; H = height of the image; W = width of the image; h = height of the search window; w = width of the search window



button, located at the lower right corner of the screen. This button is used to activate the grid and to allow the switching of the buttons between the *enabled* and *disabled* states. The patient can thus choose to enable the buttons, and then commands the system to execute his or her desired choice. Upon the execution of the function, all buttons are automatically disabled, thus recovering the initial state of the program.

It is worth noting that the system currently consists of 24 pictograms stored in an expandable database. These pictograms are further classified into the following categories:

i. **Basic Needs:** *I Am Thirsty, I Am Hungry, I Want To Go To The Toilet, I Want To Take A Shower, I Want To Sleep, I Want To Change My Clothes, I Want To Go Out, I Want To Go Home*.
ii. **Common Sentences:** *Hello!, How Are You?, I Am Fine, I Want To Watch TV, Thank You, Mom!, Dad!, I Want To Listen To Music*.
iii. **Irritation Expressions:** *I Am Feeling Cold*, *I Am Feeling Hot*, *I Have A Headache*, *I Am Feeling Pain, It Is Too Loud, Turn Off The Light, Turn On The Light, I Don't Like The Food*.

Figure 5 shows a sample of the GUI that is used by the patient.

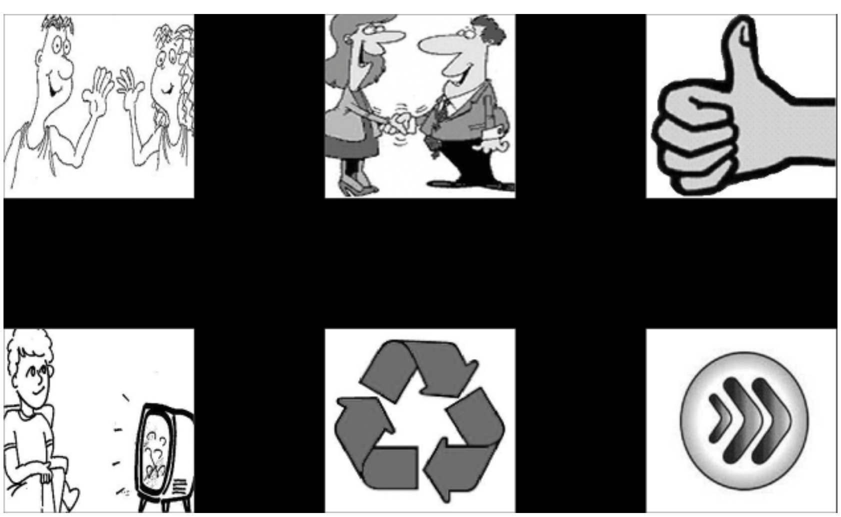
Calibration Process

A calibration procedure, upon which the point of gaze will be determined, must be executed by the patient prior to putting the system into operation. The algorithm sequentially displays a red disc, contrasted by a black background, through six predefined positions. The locked-in patient is instructed to remain focused on that disc. The center of the disc corresponds to the geometric center of the pictogram that will be displayed at that specific location. The disc is 120 pixels in diameter, and remains in a predetermined position for five seconds before moving to the next location. Eight samples are acquired at each location. The mean coordinates of the point of gaze for each disc location are subsequently determined, and will be referenced for a particular pictogram. The outcome of this procedure is a numerical grid of six rectangular coordinates representing the six different calibration points on the screen that correlate with the six pictograms per GUI. Figure 6 shows an example of the samples taken during the calibration process.

System Validation

In order to determine the overall accuracy and precision of the system, it was tested on 10 healthy adult volunteers (seven males and three females) ranging in age from 21 to 51 years (mean ± S.D. = 27.0 ± 9.5 years). Test subjects were free from any musculoskeletal or neurological disorders, had intact ocular motor

Figure 5. A sample GUI encapsulating six pictograms



control, and proper conjugate eye movement. Four of these subjects have myopia.

The validation protocol consisted of performing a series of three trials per subject in a well-illuminated environment. Test subjects were asked to execute a sequence of pre-instructed tasks, designed to test all six button positions in various combinations. Each task consisted of activating a certain pictogram by gazing at it for three seconds. The time allocated to activate a pictogram was set for 15 seconds. Beyond that time limit, the subject was considered to have failed to activate the desired pictogram and a *"miss"* was recorded for that task. The percentage of *"hits"* vs. *"misses"* was computed at the end of each trial. The values acquired from each subject during the three trials were then averaged, in order to obtain the overall system accuracy. The percentages of *"hits"* vs. *"misses"* obtained from the three trials of each subject were also used to calculate the overall system repeatability or precision.

Results

Validation results revealed a system accuracy of 96.11 ± 5.58%, and system repeatability of 94.44 ± 2.51%. These values represent the ensemble mean ± S.D. Figure 7 shows intersubject test results in terms of the rate of the *"hits"* to that of the "*misses*." Interestingly, all four myopic-test subjects exhibited reduced performance. This observation suggests that poor eye-sightedness may affect system performance, and thus requires further investigation.

Figure 6. a) An example of the samples taken during the calibration process; b) The mean values of the calibration coordinates in (a); each of the six computed markers is used to reference the corresponding pictogram on the GUI

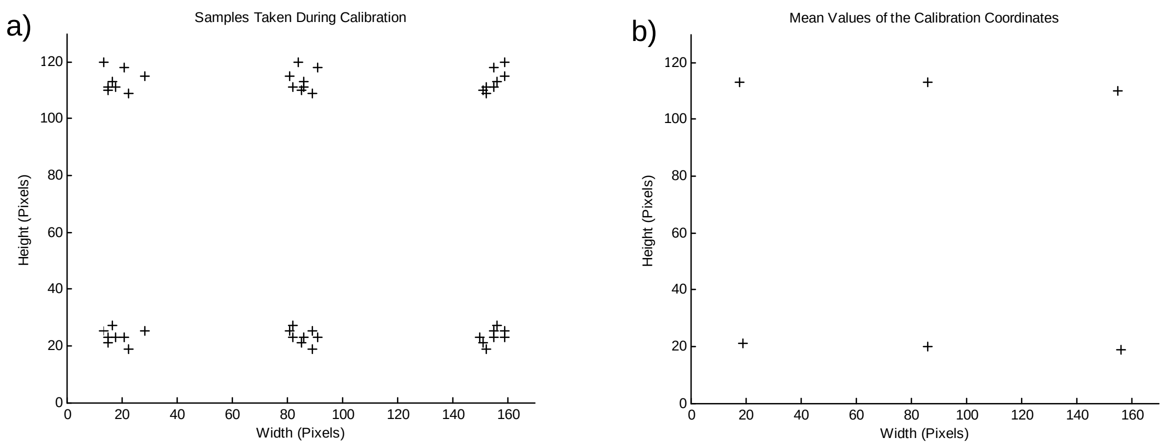
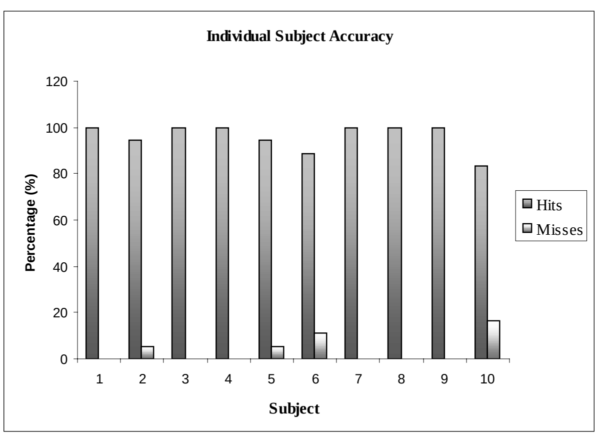


Figure 7. Intersubject accuracy (n = 10 subjects)



Discussion

In the process of developing this system, critical questions were raised and, hence, highlighted some challenging constraints that were progressively solved:

1. The choice to use a head-mounted camera was based on three key disadvantages of remotely fixed cameras. First, remotely-placed cameras should have high resolution in order to get a sharp image of the pupil. This requirement would result in increased processing resources, and hence, would worsen system complexity. Second, to obtain the position of the pupil with a remotely placed camera, three-dimensional coordinates are required; while the position of head-mounted camera relative to the eye is always constant, thereby reducing the number of spatial rectangular coordinates to two. Last, once the image of the eye falls outside the field of view of the remote camera, re-tracking the pupil would be difficult.
2. A semicircular headrest, mounted at the back of the wheelchair, was used to reduce the effect of uncontrolled head movements relative to the screen. The headrest was ergonomically designed to avoid any discomfort to the patient and, in turn, decrease the degrees-of-freedom in head motion.
3. Images captured by the camera during an eye blink result in correlation values that do not exceed a certain preset threshold, since areas of dark pixel concentration (i.e., the pupil) are no longer present. Such images are irrelevant for the calculation of the pupil position, and are therefore discarded by the system.
4. The system disables all buttons, during and following the execution of a certain function, to prevent the patient from unintentionally activating another function. Erroneous executions are avoided by adding a confirmation step, whereby the patient is requested to assert the execution of the desired command. If within 15 seconds no confirmation is received, the system aborts the execution, returns to the home page, and disables all buttons.
5. Locked-in patients are, in some cases, expected to show signs of limb spasticity. In such situations, it has been recommended by consulted rehabilitative practitioners that the arms and legs be constrained to the wheelchair in an extended, comfortable position.

Finally, it is important to mention that the system must be further tested with a population of completely locked-in patients before it can be put into clinical use.

FUTURE TRENDS

The system can be considered as a threshold of future developments. It can be modified to be implemented on an automated wheelchair where the patient would control the movement of the wheelchair by simply gazing at directional buttons on the screen. Besides, expression-building add-ins can be developed to allow the completely locked-in patient to express his or her thoughts and personality. Furthermore, the GUI could be customized to personal needs, and the database modified to include recorded statements that fit the gender, age group, and language of the particular patient. Lastly, the system can be improved by being developed as a stand-alone application.

CONCLUSION

Brain-computer interfaces are imperative in the rehabilitation of individuals with neuromuscular disorders. They hold the promise to restore independence and mobility to the completely locked-in patient. The rehabilitative system developed in this study offers the locked-in patient, of any social class, the ability of a simple yet effective communication. The advantages of this system over existing systems are low cost, low processing power, ease of operation, little training requirements, minimal disturbance to the patient, and ease of customization to any mother tongue.

ACKNOWLEDGMENT

This work was supported in part by funds from the Research Council of the American University of Science and Technology (AUST, Beirut, Lebanon).

The authors thank the team at SESOBEL (Ain el Rihani, Kesrouan, Lebanon) for their assistance; Dr. Sami Sheeshia in the Department of Computer and

Communications Engineering at AUST for his technical support, and Mrs. Henriette Skaff in the Department of Languages and Translation at AUST for editing this article.

REFERENCES

Abu-Faraj, Z.O., Mashaalany, M.J., Bou Sleiman, H.C., Heneine, J.L.D., & Katergi, W.M. (2006). Design and development of a low-cost eye tracking system for the rehabilitation of the completely locked-in patient. In *Proceedings of the 28th IEEE EMBS Annual International Conference* (pp. 4905–4908). New York, USA.

Amarnag, S., Kumaran, R.S., & Gowdy, J.N. (2003). Real time eye tracking for human computer interfaces. In *Proceedings of the 2003 International Conference on Multimedia and Expo,* (pp. III-557-560).

Barea, R., Boquete, L., Mazo, M., & López, E. (2002). System for assisted mobility using eye movements based on electrooculography. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 10*(4), 209–218.

Betke, M., Gips, J., & Fleming, P. (2002). The camera mouse: Visual tracking of body features to provide computer access for people with severe disabilities. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 10*(1), 1–10.

Böhme, M., & Barth, E. (2005). Challenges in single-camera remote eye tracking. In *Proceedings of the First Conference on Communication by Gaze Interaction* (3 pp.). Copenhagen, Denmark.

Chen, Y.L. (2001). Application of tilt sensors in human-computer mouse interface for people with disabilities. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 9*(3), 289–294.

Chen, Y.L., Tang, F.T., Chang, W.H., Wong, M.K., Shih, Y.Y., & Kuo, T.S. (1999). The new design of an infrared-controlled human-computer interface for the disabled. *IEEE Transactions on Rehabilitation Engineering, 7*(4), 474–481.

Clark, J.W. (1998). The origin of biopotentials. In J.G. Webster (Ed.), *Medical instrumentation: Application and design, third edition* (pp. 121–182). New York, NY: John Wiley & Sons, Inc.

Cui, J., Ma, L., Wang, Y., Tan, T., & Sun, Z. (2004). An appearance-based method for iris detection. In *Proceedings of the Asian Conference on Computer Vision,* (vol. 2, pp. 1091-1096).

Kennedy, P.R., & Adams, K.D. (2003). A decision tree for brain-computer interface devices. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 11*(2), 148–150.

Lee, W.J., & Galiana, H.L. (2005). An internally switched model of ocular tracking with prediction. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 13*(2), 186–193.

Li, D., Winfield, D., & Parkhurst, D.J. (2005). Starburst: A hybrid algorithm for video-based eye tracking combining feature-based and model-based approaches. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* (8 pp.).

Nagata, K., Yamada, M., & Magatani, K. (2004). Development of the assist system to operate a computer for the disabled using multichannel surface EMG. In *Proceedings of the 26th IEEE EMBS Annual International Conference* (pp. 4952–4955). San Francisco, USA.

Pfurtscheller G., Müller-Putz G.R., Pfurtscheller J., & Rupp R. (2005). EEG-based asynchronous BCI controls functional electric stimulation in a tetraplegic patient. *EURASIP Journal on Applied Signal Processing, 19,* 3152–3155.

Yu, L.H., & Eizenman, M. (2004). A new methodology for determining point-of-gaze in head-mounted eye tracking systems. *IEEE Transactions on Biomedical Engineering, 51*(10), 1765–1773.

KEY TERMS

Aphasia: A deficit in language function caused by damage to the Broca's area, Wernicke's area, or to their interconnection.

Apraxia: The inability to control voluntary movement.

Brain-Computer Interface: Software or hardware (or a combination of both) systems that allow a person to send commands to a computer program via nonordinary input devices.

Cerebral Aneurysm: A permanent abnormal blood-filled dilation of a blood vessel occurring at a flaw or weak point in the wall of an artery that supplies blood to the brain.

Cerebral Palsy: A static encephalopathy representing a group of nonprogressive neuromuscular conditions caused by injury to the immature central nervous system during its early stages of development: fetal, perinatal, and infantile.

Completely Locked-In Patient: An individual who has lost all types of motor control and communication ability with the environment, while still maintaining a certain level of cognitive functions.

Expression-Building Software: A computer program used to build small sentences by selecting one letter (or word) at a time.

Eye Tracking: The process of measuring or recording the movements of the eye.

Rehabilitation Engineering: Application of science and technology to ameliorate the handicaps of individuals with disabilities.

Stroke: It usually occurs when the blood supply to a certain part of the brain is suddenly interrupted, or when a blood vessel in the brain bursts, spilling blood into the spaces surrounding brain cells. Accordingly, brain cells die when they no longer receive oxygen and nutrients from the blood.

Traumatic Brain Injury: It occurs when a sudden trauma causes damage to the brain.

Relationship Between Performance Error and Human Information Processing

Mitsuhiko Karashima
Tokai University, Japan

INTRODUCTION

Human information processing (HIP) performance using the working memory can be assessed by two types of indicators when an HIP task is carried out. One is error occurrence, and the other is HIP time taken when the HIP task is carried out using the working memory. Errors are classified into the error caused by the task requirement exceeding some human limitation, or the error caused by carelessness, even though all human limitations still allow enough capacity to do the task (Reason, 1990). The former is regarded as an error that is caused by the lack of the HIP ability in order to do the required information processing. The latter is regarded as an error that is caused by the temporary reduction of some HIP ability such as attention. Even though there are many kinds of factors of error generation, from the view point of HIP, error can be considered to be caused by the relationship between the required quantity or quality of the information processing and the HIP ability. The characteristics of HIP can be considered to influence error generation directly. In this chapter the characteristics of HIP related to the error are illustrated with the results of the experiments (Karashima, Okamura & Saito, 1994; Karashima & Saito, 2001).

BACKGROUND

In regard to error occurrence, the difference of the error ratio (calculated by the ratio of the number of error occurrences divided by the number of trials) in the HIP tasks has been explained by the degree of the efficiency of the information processing and storage with the interference in the working memory (Carpenter, Just, Keller, Eddy & Thulborv, 1999; Daneman & Carpenter, 1980, 1983; Just & Carpenter, 1992). Interference in the working memory means that the working memory is a system responsible for processing and storing information, and the processing and storage compete with the limited-capacity workspace in the working memory (Baddeley & Hitch, 1974; Carpenter et al., 1999; Case, Kurland & Goldberg, 1982; Danemen & Carpenter, 1980, 1983; Miyake, Just & Carpenter, 1994). These suggest that the error ratio increases when the storage information increases, and that the ways of the increase are different between the information processing in the HIP tasks.

On the other hand, the error ratio in the HIP task is rarely 0%. The HIP task is performed correctly in some cases but in some others performed incorrectly, even though all the conditions and the protocol for each subject are fixed. This means that there might be the variation of the performance under the same task. In a typical case, the subjects try to memorize and recall 7±2chunk, which is reported as the upper limit by Miller (1956) (Broadbent, 1975). But it is difficult to explain the variation by the influence of the working memory resource. The variation has not been studied enough to explain these characteristics of the error occurrence.

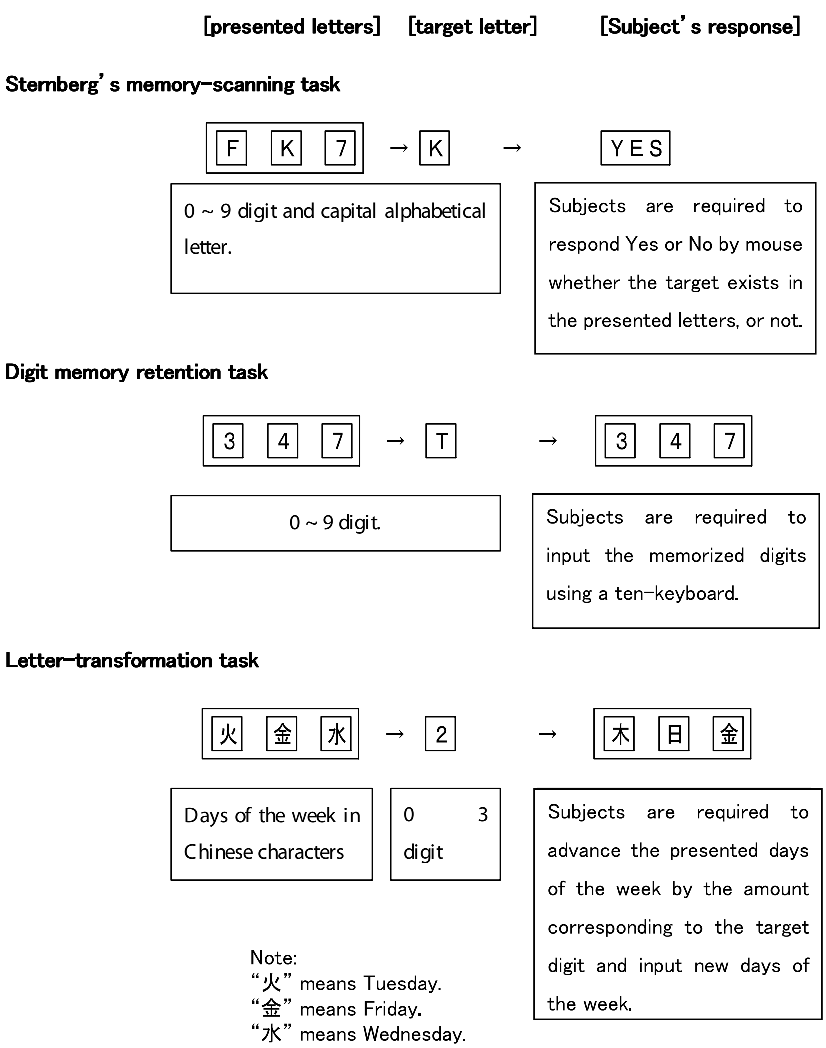
MAIN FOCUS OF THE CHAPTER

The Transformation of the Human Information Processing Ability by the Content of Human Information Processing

Generally, error occurs when the HIP ability is not enough to do the required information processing. This ability is sometimes transformed by the content of the required information processing or the environments. As the illustration of these characteristics of HIP, it is known that the HIP ability to deal with the presented information quantity is influenced by the constraints of the working memory, the interference of the working memory caused by the constraints, and the trade-off caused by the interference between information processing and storage. The constraints mean that the working memory resource has the capacity; the interference means that the information processing and storage compete with the capacity. The trade-off means that the capacity of information storage in the working memory decreases when the difficulty of information processing in the working memory increases, while the information processing in the working memory becomes more delayed when the quantity of information storage in the working memory increases. Because of the trade-off, concretely, the information processing ability to deal with the presented information quantity decreases as the difficulty of information processing in the working memory increases.

Karashima, et al. (1994) examined the relationship between the presented information quantity and the error ratio in three difficulties' tasks: Sternberg's memory-scanning task (Sternberg, 1966), digit memory retention task, and letter-transformation task. Figure 1 shows the details of three tasks. The Sternberg task is easiest, and the letter-transformation task is most difficult in three tasks. His experimental results reveal that the information processing ability to deal with the presented information quantity decreases as the difficulty of HIP increases. As Figure 2 shows, for example, in the case of six presented letters, the error ratio of letter-transformation task is more than 50%, although the error ratios of the other tasks are less than 10%.

Figure 1. Three experimental tasks



R

The Fluctuation of Human Performance by the Fluctuation of the Human Information Processing Ability

Generally, error doesn't occur when the HIP ability is enough to do the required information processing. However, the careless or dull condition causes error such as slip in this situation. This kind of error is interpreted by the lack of concentration. On the other hand, from the viewpoint of HIP, this error can be interpreted by the phenomena that the HIP ability decreases temporarily to less than the sufficiency ability to do the required information processing. This temporary decrease of the HIP ability can be interpreted by the fluctuation of the HIP ability itself.

Karashima and Saito (2001) examined the time series of the error occurrences in the HIP task. Figure 3 shows the experimental task and the required protocol of the subjects. The spectral slope of the time series of the error occurrences was similar to the spectral slope of 1/f fluctuation, which was between -0.5 and -1.5 (Voss & Clarke, 1978) and frequently appeared in the spectral densities of the time series in some phenomena (Johnson, 1925; Mandelbrot & Wallis, 1969; Musha & Higuchi, 1977), some classical music (Voss & Clarke, 1978), and some human functions such as heartbeat period (Kobayashi & Musha, 1982), alpha wave in

Figure 2. Mean error ratio in three tasks (M±SD, n=13)

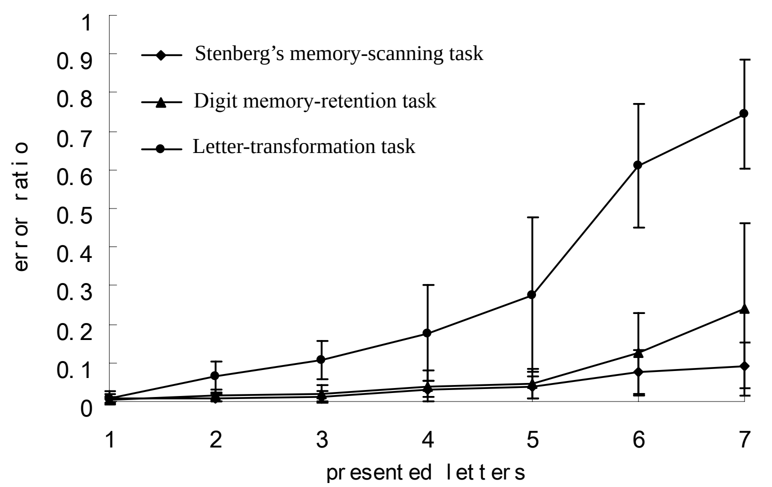
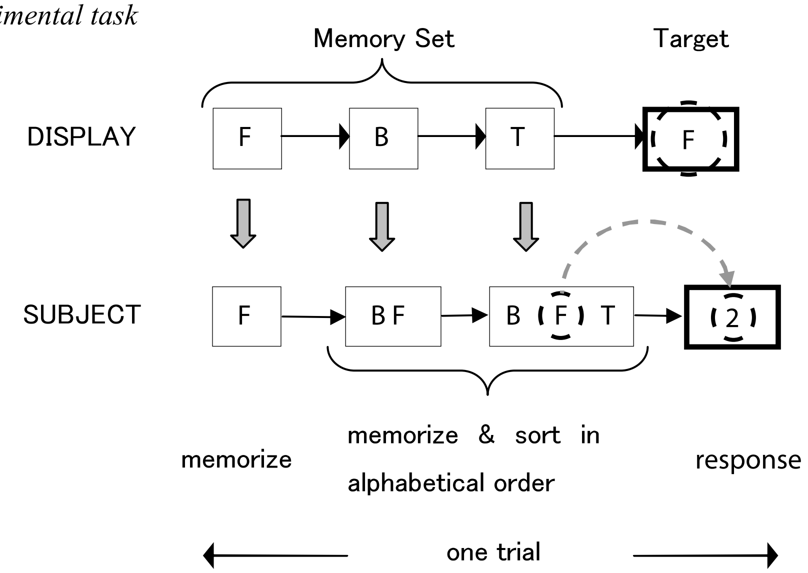


Figure 3. The experimental task



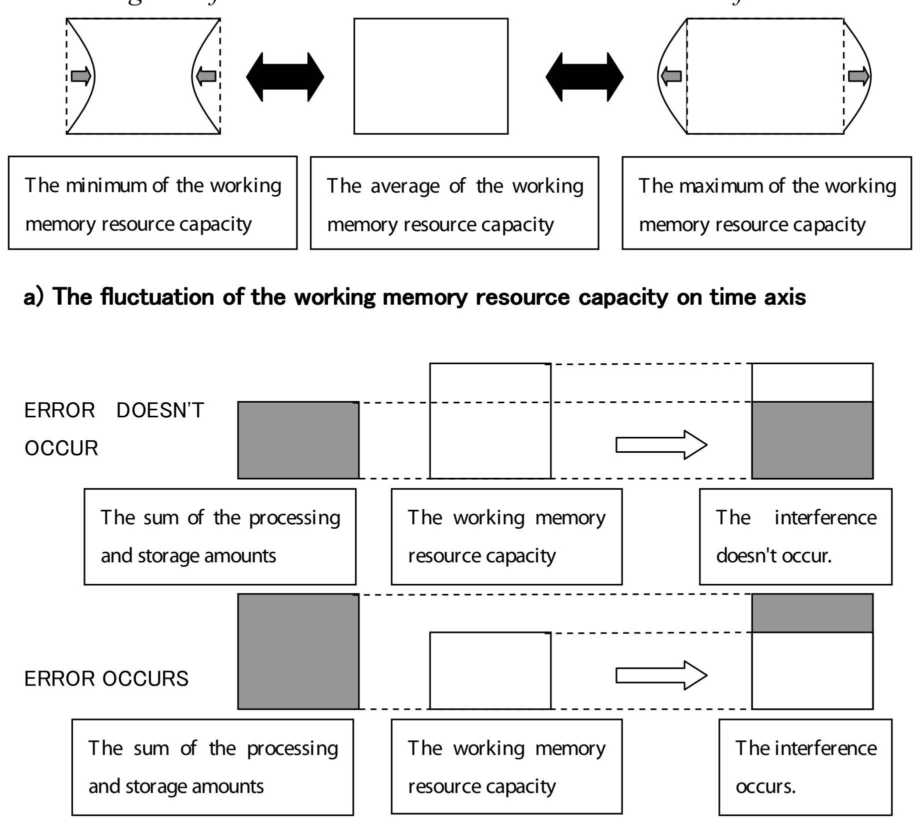
the brain (Suzuki, Ofaka, Kosugi, Ikebe, Matsuoka & Takakura, 1980), human cognitive functions (Gilden, Thornton & Mallon, 1995; Musha, Katsurai & Teramachi, 1985), and so on. It reveals that the time series of the error occurrences has some particular fluctuation in the same information processing task and the stable experimental conditions. It suggests that the HIP ability might have the fluctuation. The fluctuation of the HIP ability can be the main reason why errors sometimes occur when the HIP ability is enough to do the required information processing.

FUTURE TRENDS

The variation of the error ratio in the HIP tasks could be explained by the concept of the working memory resource capacity in the conventional working memory models proposed by Baddeley (1986) or Just & Carpenter (1992). However, it is difficult to explain the variation of the error occurrence where the protocol is fixed by the precedent concept, because the variation of the error occurrence implies the contradictive facts that the interference in the working memory occurs in some cases, while in some others, the interference doesn't occur and that the degree of the efficiency changes even though the sum of the processing and storage amounts is fixed under the same task. In addition, the characteristic fluctuation in the time series of the error occurrence also cannot be explained by the precedent concept.

It might be possible that these contradictions are due to the constant working memory resource capacity in the conventional working memory models. Karashima & Saito (2001) proposed the new theoretical hypothesis on the working memory resource capacity, which could explain the variation of the error occurrence without any contradiction in terms of the interference in the

Figure 4. FLOWM and diagram of the case in which error doesn't occur and of the case in which error occurs



a) The fluctuation of the working memory resource capacity on time axis

b) The relation among the sum of the processing and storage amounts, the working memory resource capacity, and the interference of the processing and storage in the working memory

working memory. This new hypothesis was named FLOWM (FLuctuation Of Working Memory resource capacity). FLOWM adds the concept of a fluctuation on time axis to the concept of the working memory resource capacity in the conventional working memory models, as shown in Figure 4. It is defined by FLOWM that the working memory resource capacity varies with some fluctuation on time axis. It is also defined by FLOWM that the interference occurs when the working memory resource capacity is smaller than the sum of the processing and storage amounts in the working memory and the interference doesn't occur when the capacity is larger, and that the interference causes the error occurrence. The variation of the error occurrence can be explained by applying FLOWM. In some cases, the error occurs because the working memory resource capacity is smaller than the sum of the processing and storage amounts, and the interference occurs. In some others, the error doesn't occur because the capacity is larger than the sum of the amounts, and the interference doesn't occur even though the protocol and the sum of the amounts are fixed. Figure 5 shows the explanation of the variation of the error occurrence where the error ratio is 50%. FLOWM can explain the variation of the error ratio in the same way as the precedent concept in the conventional model.

By applying FLOWM, the time series of the error occurrence can be regarded as the time series of the cases in which the sum of the processing and storage amounts becomes larger than the fluctuating capacity of the working memory resource when the task is carried out, thereby leading to the possibility that the characteristic fluctuation in the time series of the error occurrence is produced by the fluctuation of the working memory resource capacity.

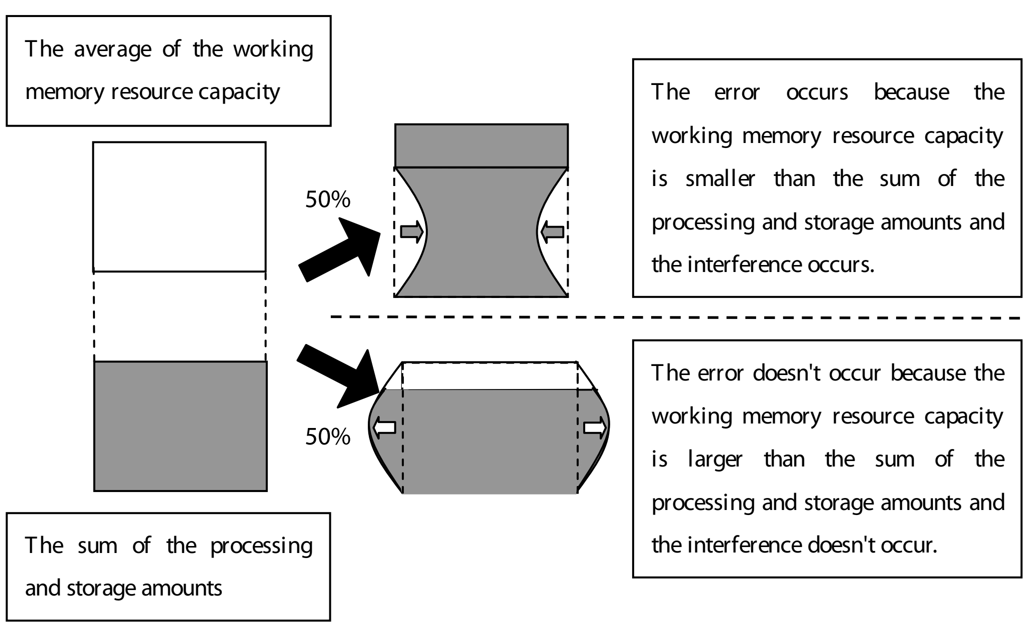
Karashima & Saito (2001) proposed the new theoretical hypothesis, FLOWM, on the working memory resource capacity, which could explain the variation of the error occurrence without any contradiction in terms of the interference in the working memory. In his paper, however, the validity of FLOWM was not directly confirmed because any physiological indicators of the human brain activities were not measured. Only the performances were measured. Further studies are left for confirming the validity of FLOWM from the neurophysiological standpoint.

CONCLUSION

In this chapter, the characteristics of HIP related to the error were illustrated:

1. The information processing ability to deal with the presented information quantity decreases as the difficulty of HIP in the task increases, because of

Figure 5. Explanation of the variation of error occurrence where the error ratio is 50% by applying FLOWM



the trade-off caused by the interference between information processing and storage in the working memory.

2. Errors sometimes occur when the HIP ability is enough to do the required information processing because the HIP ability might have the fluctuation.

FLOWM, the new theoretical hypothesis on the working memory resource capacity that could explain the characteristics without any contradiction in terms of the interference in the working memory, was introduced.

REFERENCES

Baddeley, A.D. (1986). *Working memory*. New York: Oxford University Press.

Baddeley, A.D., & Hitch, G.J. (1974). Working memory. In G. Bower (Ed.), *The psychology of learning and motivation* (pp. 47–90), New York: Academic Press.

Broadbent, D.E. (1975). The magic number seven after fifteen years. In A. Kennedy, & A. Wilkes (Eds.), *Studies in long term memory* (pp. 3–18). London: Wiley.

Carpenter, P.A., Just, M.A., Keller, T.A., Eddy, W., & Thulborn, K. (1999). Graded functional activation in the visuospatial system with the amount of task demand *Journal of Cognitive Neuroscience, 11*, 9–24.

Case, R., Kurland, M., & Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. *Journal of Experimental Psychology, 33*, 386–404.

Daneman, M., & Carpenter, P.A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19*, 450–466.

Daneman, M., & Carpenter, P.A. (1983). Individual difference in integrating information between and within sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9*, 561–584.

Gilden, D.L., Thornton, T., & Mallon, M.W. (1995). 1/f noise in human cognition. *Science, 267*, 1837–1839.

Johnson, J.B. (1925). The Schottky effect in low frequency circuits. *Physical Review, 26*, 71–85.

Just, M.A., & Carpenter, P.A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99*(1), 122–149.

Karashima, M., Okamura, T., & Saito, M. (1994). A study on the quantity of information presentation for the work supported by short-term memory, *The Japanese Journal of Ergonomics, 30*(2), 63–69.

Karashima, M., & Saito, M. (2001). A study on the error occurrence and the human information processing time influenced by the fluctuation of the working memory resource capacity. *International Journal of Cognitive Ergonomics, 5*(2), 91–109.

Kobayashi, M., & Musha, T. (1982). 1/f fluctuation of heartbeat period. *IEEE Trans., BME-29*, 456–457.

Mandelbrot, B.B., & Wallis, J.R. (1969). Some long-run properties of geophysical records. *Water Resources Research, 5*(2), 321–340.

Miller, G.A. (1956). The magical number seven, plus or minus two. *Psychological Review, 63*, 81–97.

Miyake, A., Just, M.A., & Carpenter, P.A. (1994). Working memory constraints on the resolution of lexical ambiguity: Maintaining multiple interpretations in neutral contexts. *Journal of Memory and Language, 33*, 175–202.

Musha, T., & Higuchi, H. (1977). The 1/f fluctuation of a traffic current on an expressway. *Japan Journal of applied Physics, 15*, 1271–1275.

Musha, T., Katsurai, K., & Teramachi, Y. (1985). Fluctuations of human tapping intervals. *IEEE Trans., BME-32*, 578–582.

Reason, J. (1990). Human error. Cambridge, UK: Cambridge University Press

Sternberg, S. (1966). High-speed scanning in human memory. *Science, 153*, 652–654.

Suzuki, M., Odaka, T., Kosugi, Y., Ikebe, J., Matsuoka, K., & Takakura, K. (1980). Frequency fluctuation of human EEG and quantitative analysis of relief of pain. *Rep. Tech. Res. Group IECE Japan, MBE80-59*, 33–40.

Voss, R.F., & Clarke, J. (1978). "1/f noise" in music: Music from 1/f noise. *Journal of Acoust. Soc. Am., 63*(1), 258–263.

KEY TERMS

1/f Fluctuation: 1/f fluctuation is also called 1/f noise or pink noise. The power spectral density of this noise is proportional to the reciprocal of the frequency. It occurs in many fields. It is reported that people tend to feel comfortable to some time series that has 1/f fluctuation. The origin of its name is from being intermediate between white noise, $1/f^0$ noise, and red noise, $1/f^2$ noise.

Error Ratio: Error ratio is generally calculated by the ratio of the number of error occurrences divided by the number of trials in HIP task. Error ratio is measured as one of the indices of the performance, mental workload, and fatigue in the task.

FLOWM (*FLuctuation of Working Memory Resource Capacity*): FLOWM was proposed for explaining the performance such as the variations of the error occurrence and the HIP time, which could not be explained without contradiction by the conventional models such as Baddeley's model due to the constant working memory resource capacity. FLOWM is the new hypothesis that adds the concept of a fluctuation on time axis to the concept of the working memory resource capacity in the conventional working memory models.

Working Memory: Working memory is the theoretical framework in cognitive psychology that describes the processes for processing and storing information. Baddeley & Hitch (1974) made the model of working memory. It consists of one main system, "central executive," and two slave systems, "phonological loop" and "visuospatical sketch pad."

Working Memory Resource Capacity: Working memory consists of one main system and two slave systems. Every system has workspace for information processing and storage, and each workspace has the limited capacity. When the amount of information processing and storage in each system nears full capacity, the interference between information processing and storage occurs easily and the interference makes the performance of the information processing decrease or causes the loss of stored.

Relationship Management Competence and Organizational Performance

Murako Saito
Waseda University, Japan

INTRODUCTION

The competence of organizational management is required to cope with the complexity of technology and the diversity of social demands for maintaining good interpersonal relationships through the support of advanced technology. The competencies, such as self-management and interpersonal relationship management, play a crucial role in improving individual and organizational performance. Experiences of interpersonal relationships at work are multifaceted with value consciousness, mood states at work, and members' work attitudes, as well as technical knowledge and skills that are constituted of explicit cognitive information. Organizational climate or culture is focused on recent studies to improve human performance by avoiding erroneous actions, accidental occurrences, or withdrawal attitudes such as absenteeism and turnover intention. The participants to organization are expected to act as actors and some of them as leaders to help develop themselves and coworkers, building bonds to collaborate with team members, for improving organizational performance, and for providing a high quality of service. This article is focused on the effect of mood states at work on workers' perceived health and perceived performance, and on the effect of the competence of interpersonal relationship management on organizational performances.

BACKGROUND

Emotional competences represented by recognition of self and others' competences, emotional regulation of self and relationship management competence, and also mood states at work are important modulators in improving quality of care (QoC). We have tried to categorize care into four types by using two axes: one is the axis of care, lower or higher; the other is the axis of group level, individual or social, such as egoistic type (lower level of care, individual), bureaucratic type (lower level of care, social), specialist type (higher level of care, individual), and dynamic collaboration type (higher level of care, social). Our studies in healthcare sectors suggest that dynamic and collaboration types of care categorized by two axes play an important role in improving quality of team care and also organizational performances.

In order to identify the important role of dynamic and collaboration type of care, some evaluation indicators such as mood states at work, perceived health state, and emotional intelligence competence were applied in our studies. Our studies suggested that job consciousness, emotional states at work, and occupational health states gave critical effects on individual and organizational performance. The profile of mood states (POMS) developed by McNair, Lorr, and Droppleman (1971) was applied in investigations as to the measurement of mood states. The POMS consisted of six T-values: tension and anxiety (T-A), depression- dejection (D-D), anger-hostility (A-H), vigor (V), fatigue (F), and confusion (C). Perceived health status was measured by applying the Todai Health Index (THI) developed by Suzuki, Aoki, and Yanai, (1976) and Suzuki and Roberts (1991). Perceived health status was depicted as a radar chart constituted of 12 scale values, vague complaints (SUSY), respiratory (RESP), eye and skin (EYSK), mouth and anal (MOUT), digestive (DIGE), irritabilities (IMPU), lie scale (LISC), mental instability (MENT), depressiveness (DEPR), aggressiveness (AGGR), nervousness (NERV), and life irregularity (LIFE). In addition, with the pattern of health structure, health level was also diagnosed by using DF values (discriminant function values) for psychosomatics, neurotics, and schizophrenics, as shown in Figure 2. The larger the DF value, the greater the probability of disease. The competence of interpersonal relationship management was measured by applying Emotional Intelligence Competence (EIC), which was a learned ability and was classified into self-awareness, social awareness, self-management, and relationship management (Boyatzis, Goleman & Rhee, 2000; Cherniss

& Goleman, 2001; Goleman, Boyatzis & McKee, 2002).

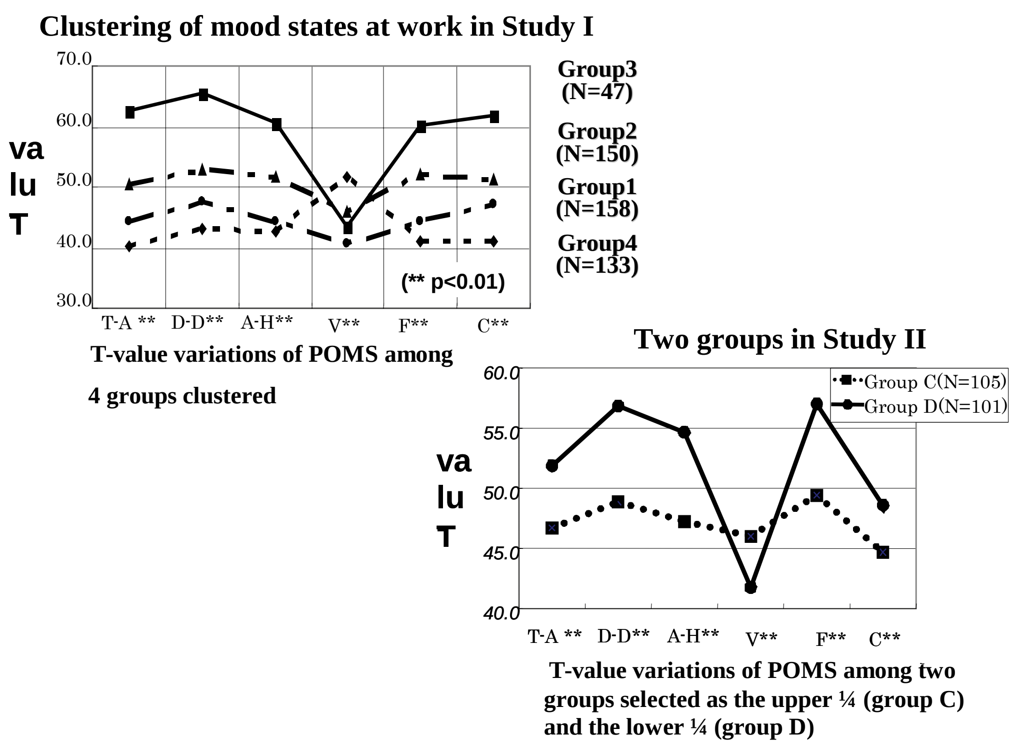
Performance reliability was measured by nine items of work environmental conditions called common performance conditions (CPCs), asking, How do you cognize the adequacy of (1) organization, (2) working conditions, (3) adequacy of Man-Machine Interface and operational support, (4) availability of procedures/plans, (5) number of simultaneous goals, (6) available time, (7) time of day/circadian rhythm, (8) adequacy of training and preparation, (9) and crew collaboration quality, which are developed by Hollanagel (1993, 1998). Subjects are requested to select one of four answers: very efficient, efficient, inefficient, and deficient. Erroneous human behaviors are predicted by measuring performance reliability both in improved reliability and reduced reliability.

MAIN FOCUS OF THE ARTICLE

The focus of this article is placed on the relation of mood states at work with workers' health conditions and with perceived performance, and also placed on the relationship of interpersonal relationship management competence with organizational performances of communication accuracy, team coherence, and performance reliability. As shown in Figure 1, subjects were classified into four groups in Study I, as shown in the upper left figure, and two groups in Study II, as shown in the lower right figure. An asymmetrical pattern of mood states at work measured by the POMS was observed between Group 3 and Group 4 in Study I.

Scores of interpersonal relationship managements were obtained by the questionnaire inventory of the EIC. The framework of the EIC consisted of four conceptual domains: self-awareness, social awareness, self-management, and relational management. Poor interpersonal relationship management in individual and collective levels (i.e., inappropriate recognition and regulation management) often causes the decrement of both task performance and contextual performance. The results shown in Figures 4 and 5 are on the comparison of organizational performances among the degrees of the EIC, communication accuracy in Figure 4, team member resources in Figure 5, and also the results in the section of performance reliability. Interpersonal relational management, mood-states at work measured

Figure 1. Clustering of mood states at work



in industrial fields, and healthcare sectors play crucial roles as moderators or mediators in changing health structure and health level as well as job performances (Murakami, Inoue, & Saito, 2007; Saito, Ikeda, & Seki, 2000; Saito, Inoue, & Seki, 2002, 2003; Saito, Murakami, & Karashima, 2007).

Classification of Mood States at Work

Asymmetrical patterns of mood states at work measured by the POMS were observed between Group 3 and Group 4 in Study I, and apparently low T-value in comparison with other five T-values in Group D in Study II were observed by clustering sample subjects of industrial workers as shown in Figure 1. T-values of vigor (V) among six T-values (T-A, D-D, A-H, V, F, and C) appeared as a characteristically lower value in Group 3 and Group D.

Health Structure and Level by Mood States at Work

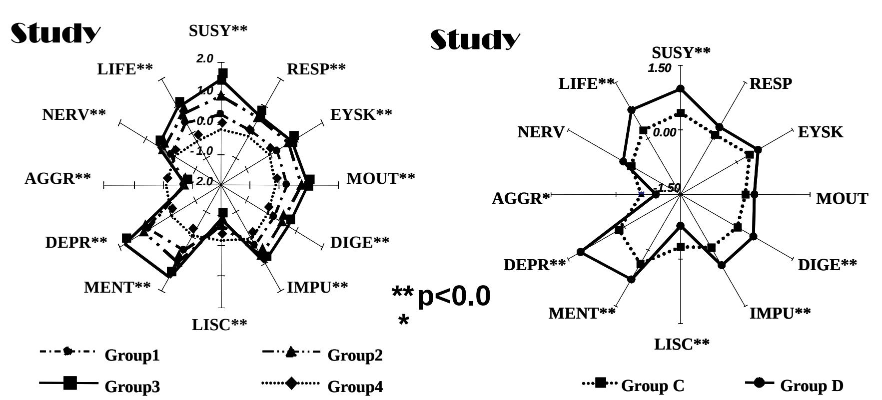
Perceived health status was depicted as a radar chart constituted of 12 scales: vague complaints (SUSY), respiratory (RESP), eye and skin (EYSK), mouth and anal (MOUT), digestive (DIGE), irritabilities (IMPU), lie scale (LISC), mental instability (MENT), depressiveness (DEPR), aggressiveness (AGGR), nervousness (NERV), and life irregularity (LIFE), as shown in Figure 2. In Group 3 in Study I and Group D in Study II, characteristically deviated patterns of health structure appeared in comparison with other groups. The least deviation in the standard health pattern of Japanese workers was seen in Group 4, of which T-value of Vigor was the highest among four groups; DF values for diagnosing psychosomatics, neurotics, and schizophrenics were the lowest among four groups; and Group 3 was the highest probability of psychosomatics, neurotics, and schizophrenics. The larger the DF value, the greater the probability of disease. In Study II, health structure of Group D was significantly deviated, especially in the scales of DEPR and MENT; and DF values in psychosomatics and neurotics were significantly higher than those of Group C. Mood states at work play decisive roles in improving perceived health.

Perceived Performance by Mood States at Work

Mood states at work also have an effect on perceived performance. Perceived performance (actuality/potentiality) in Group3 appeared the lowest among the four groups, while perceived performance in Group 4 was the highest, as shown in Figure 3.

Figure 2. Radar graph composed of 12 scales and discriminate diagnosis of perceived health among four groups in Study I and two groups in Study II

Perceived health by mood states at work



	Group1	Group2	Group3	Group4
PSD**	*-0.013*	*-0.739*	*-0.133*	*-1.695*
NEUR**	*-1.008*	*-0.751*	*0.399*	*-1.982*

	Group C	Group D
PSD**	*-1.12*	*-0.42*
NEUR**	*-1.65*	*-0.40*

Figure 3. Perceived performance among four groups

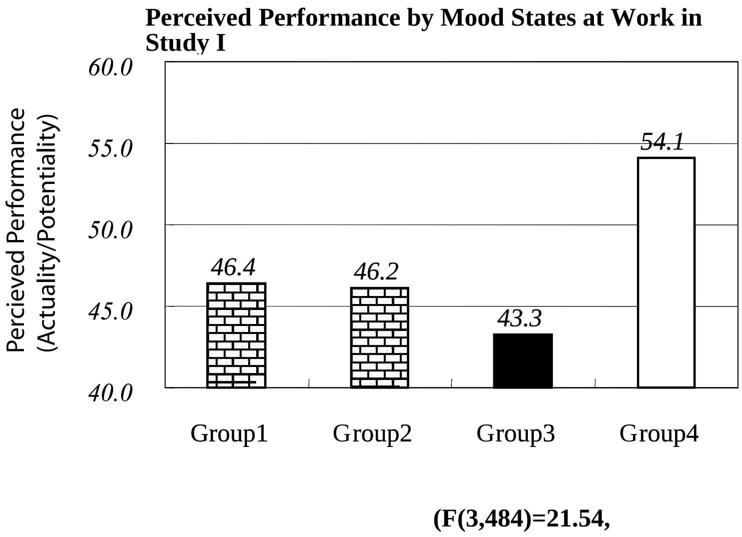
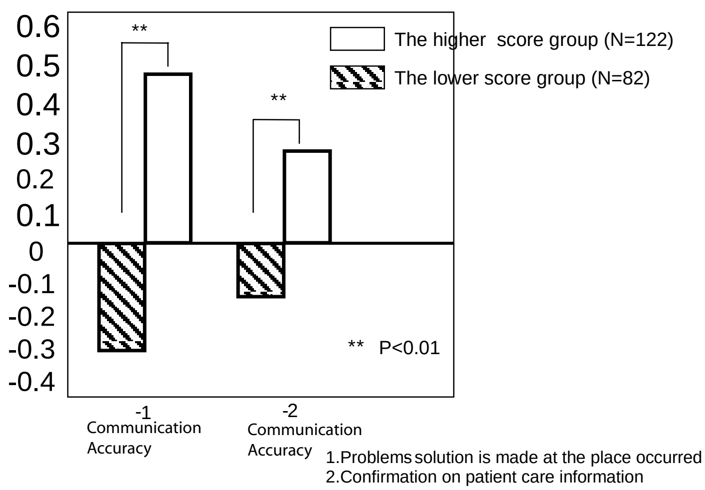


Figure 4. Comparison of communication accuracy between the higher and lower groups of interpersonal relationship management (N=124)



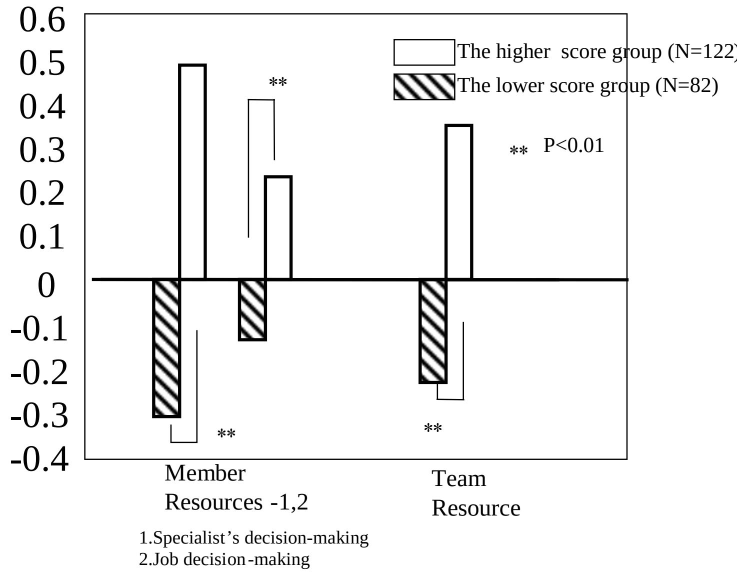
Communication Accuracy and the EIC

Recognition in social awareness and interpersonal relationship management gave strong effects on the performances, such as accuracy of communication, development of team and member resources, and performance reliability (Murakami, 2005; Saito et al., 2003, 2007). As shown in Figure 4, communication accuracy-1, which means the questions and problems are solved at the place occurred, and communication accuracy-2, which is confirmation of information on patient care and management procedures concerned, are significantly different between the higher and lower score groups in the EIC. This result suggests that the higher level of the EIC enhances accuracy of communication.

Team Member Resources and the EIC

Member resources-1 means decision-making as a specialist on the basis of established disciplinary knowledge, while member resources-2 denotes decision-making as practitioners; they were significantly different between the higher and the lower groups of

Figure 5. Comparison of member resources and team resources between the higher and lower groups of interpersonal relationship management



the EIC, as shown in Figure 5. Team resource representing effectiveness of teamwork, enhancing self-efficacy, also are significantly different between the two groups, as also shown in Figure 5. This result suggests that the EIC plays a significant role in enhancing team member resources.

Performance Reliability and the EIC

Performance reliability is measured by the CPCs and evaluated by two scales of improved reliability and reduced reliability. The higher score group of the EIC was significantly higher in improved reliability, while the lower score group of the EIC was significantly higher in reduced reliability. Our study results suggest that emotional intelligence competence plays a significant role in improving performance reliability, which is one of the important indicators of predicting human error (Saito, Murakami, & Karashima, 2007).

Relational management competence has a significant effect on performance reliability of care-providers through communication accuracy among team members and other professional staff. The results suggested that the difference of care performance in healthcare sectors and of work performance in industry could be explained by measuring interpersonal relationship management competence, as shown in Figure 4 and Figure 5. Intangible human assets embedded in the organization, such as interpersonal relationship management competence, play a great role in improving care performance reliability and other organizational performances.

LIMITATIONS AND FUTURE TRENDS

Empirical evidence in this article is limited in subject number, investigation area, and also in the method of measuring and analyzing. Continuing surveys in the future are required to testify to the relation of organizational performance with interpersonal relationship management by the use of contextual factors of organizational environment in addition to communication accuracy, team reciprocity, and performance reliability as shown in this article. It is of importance in redesigning organization to predict organizational performance in terms of a service provider's competence on interpersonal relationship management and its perceived health status. The enhancement of work ability and management competence of employees necessitates prompt and appropriate adapting to the changing technological and organizational environments, and improving the QOC, especially dynamic collaboration type of care, which promises to gain intangible human assets.

CONCLUSION

1. Mood states at work, emotional intelligence competences that were used to evaluate interpersonal relationship management, have a significant effect on organizational performances, such as perceived performance, communication accuracy, team member resources, and performance reliability both in improved and reduced reliability. These results suggest that emotional and mood regulation competences play a crucial role in improving performance both on individual and collective levels.
2. Studies shown in this article are limited in evaluating organizational performance and also in collecting the subjects in various working situations. Further studies were left for generalization of the results shown in this article.

REFERENCES

Beer, S. (1979). *The heart of enterprise*. John Wiley and Sons, Inc.

Boyatzis, R.E., Goleman, D., & Rhee, K. (2000). Clustering competence in emotional intelligence: Insight from the emotional intelligence inventory. In R. Baron & J.D.A. Parker (Eds.), *The handbook of emotional intelligence: Theory, development, assessment and application at home, school and in the workplace* (pp. 343–362), Jossey-Bass.

Boyatzis, R.E., & McKee, A. (2005). *Resonant leadership*. Harvard Business School Press.

Cherniss, C., & Goleman, D. (2001). *The emotionally intelligent workplace: How to select for, measure, and improve emotional intelligence in individuals, groups and organizations*. Jossey Bass.

Espejo, R., Schumann, W., Schwaninger, M., & Bilello, U. (1996). *Organizational transformation and learning: A cybernetic approach to management*. John Wiley and Sons, Inc.

Flood, R. (1993). *Beyond TQM*. John Wiley and Sons, Inc.

George, J.M., & Jones, G.R. (1996). The experience of work and turnover intentions: Interactive effects of value attainment, job satisfaction and positive mood. *Journal of Applied Psychology, 81*(3), 318–325.

George, J.M., & Jones, G.R. (1997). Experiencing work: Values, attitudes and moods. *Human Relations, 50*(4), 393–416.

Goleman, D., Boyatzis, R.E., & McKee, A. (2002). *Primal leadership: Realizing the power of emotional intelligence*. Harvard Business School Press.

Hochwarter, W.A., Perrewe, P.L., & Ferris, G.R. (1999). Job satisfaction and performance: The moderating effects of value attainment and affective disposition. *Journal of Vocational Behavior, 54*, 296–313.

Hollnagel, E. (1993). *Human reliability analysis: Contxy and control*. London, Academic Press.

Hollnagel, E. (1998). *Cognitive reliability and error analysis method: CREAM*. Elsevier.

McNair, D.M., Lorr, M., & Droppleman, L.F. (1971). *Manual for the profile of mood states (POMS)*. San Diego, CA: Educational and Industrial Testing Service.

Murakami, G. (2005). *A study on care cognition and performance reliability in team nursing.* Dissertation abstracts for master's degree, Waseda University (pp. 245-248).

Murakami, G., Inoue, T., & Saito, M. (2007). On cognitive reliability of hospital organizational environment and work performance. *Int J Healthcare Technology and Management*, *8*(3-4), 388-398.

Reason, J. (1997). *Managing the risk of organizational accidents*. Ashgate.

Saito, M., Ikeda, M., & Seki, H. (2000). *A study on adaptive action development: Interference of mood states with perceived performance and perceived efficacy*. Proceedings of the IEA and HFES Congress, San Diego, California, Vol. 5, 48–51.

Saito, M., Inoue, T., & Seki, H. (2002). *The effect of mood states on perceived performance and on perceived health*. Proceedings of the 7th International Congress on Behavioral Medicine, Helsinki, Finland.

Saito, M., Inoue, T., & Seki, H. (2003). *Organizational control and performance reliability in hospital nurses*. Proceedings of the IEA and HHEF Congress, T-286.

Saito, M., Murakami, G., & Karashima, M. (2007). Effect of communication and emotional types on team reciprocity in healthcare organization. *International Journal of Health Care Technology and Management, 8*(3-4), 196-208.

Schwaninger, M. (2000). Managing complexity: The path toward intelligent organization. *Systemic Practice and Action Research, 13*(2), 207–241.

Seers, A. (1989). Team member exchange quality: A new construct for role-making research. *Organizational Behavior and Human Decision Processes, 43,* 118–135.

Seers, A., Petty, M.M., & Cashman, J.F. (1995). Team member exchange under team and traditional management: A naturally occurring quasi-experiment. *Group & Organizational Management, 20*, 18–38.

Suzuki, S., Aoki, S., & Yanai, H. (1976). An introduction to a new health questionnaire, the Todai health index. *Progress in Medicine (Igaku-no-ayumi), 99*, 217–225 (in Japanese).

Suzuki, S., & Roberts, R.E. (1991). *Methods and applications in mental health surveys: The Todai health index*. University of Tokyo Press.

KEY TERMS

Dynamic Collaboration Type of Care: Assessment of the quality of care has been inquired, but we thought that classification of care is required before assessment. We made a classification by using two axes: the axis of care, higher or lower; and the axis of care-provider, individual or group/social. We found four types of care: egoistic, bureaucratic, specialist, and dynamic collaborate. Our study results suggest that dynamic collaborate type of care plays an important role in improving quality of team care. This type of care leverages organizational performance.

Perceived Health: The measurement of the pattern of perceived health was made by using 12 scales of vague complaints (SUSY), respiratory (RESP), eye and skin (EYSK), mouth and anal (MOUT), digestive (DIGE), irritabilities (IMPU), lie scale (LISC), mental instability (MENT), depressiveness (DEPR), aggressiveness (AGGR), nervousness (NERV), life irregularity (LIFE), and a health level also made by using DF (discriminate function) values for psychosomatics and neurotics, which were developed by Suzuki, et al. (1976).

Perceived Performance: Perceived performance used in this article was measured by the use of methodology developed by Beer (1979) and applied by many other researchers such as Flood (1993), Espejo, Schumann, Schwaninger & Bilello (1996), Schwaninger (2000), and their co-workers. Perceived performance = Actuality/Potentiality, whereas actuality means actual work dealing with daily working time; potentiality is the best possible work when all the constraints in work environment are removed and you think it is your potential resource.

Performance Reliability: Developed by Hollnagel (1998) in predicting human error during work, measured by using nine common performance conditions (CPCs) with a four-point scale: very efficient, efficient, inefficient, and deficient. Improved reliability and reduced reliability are also measurable by using CPCs.

Team Resources: Team resources were measured by applying the TMX (Team Member Exchange Quality) developed by Seers (1989). Team resource was evaluated in this article to define mutual support, group cohesiveness, trust, responsibility for team goals, and team reciprocity, as defined by the developer.

R

The Relevance of Computer-Aided-Diagnosis Systems in Microscopy Applications to Medicine and Biology

Paolo Soda
Università Campus Bio-Medico di Roma, Italy

Giulio Iannello
Università Campus Bio-Medico di Roma, Italy

INTRODUCTION

Microscopy is a branch of medical sciences strictly related to image analysis and interpretation. Various sectors of medicine make use of micrographs[1] for different purposes. Moreover, since different biological substrates can be used, micrographs are usually divided into cytological and tissue images.

Humans are limited in their ability to distinguish similar objects, and to diagnose diseases during image interpretation because of noise and of their nonsystematic search patterns. In addition, the vast amount of image data generated by imaging devices makes the detection of potential diseases a burdensome task and may cause oversight errors. Developments in computer vision and artificial intelligence in medical imaging have shown that computer-based systems can be employed in different fields of medicine and biology, especially in those based on digital images (e.g., radiography, mammography, and echography, to name a few).

In microscopy, many qualitative or semiquantitative features have to be analyzed, demanding for qualified personnel that are not always available. Moreover, in several microscope applications the lack of quantitative information limits their reproducibility. For these reasons, in the last years the main research efforts on micrographs have been directed toward automated image processing and analysis, particularly in the field of both image segmentation and classification. The former aims at subdividing an image into its constituent regions or objects, whereas the latter works toward the recognition of image parts or objects detected by the segmentation process.

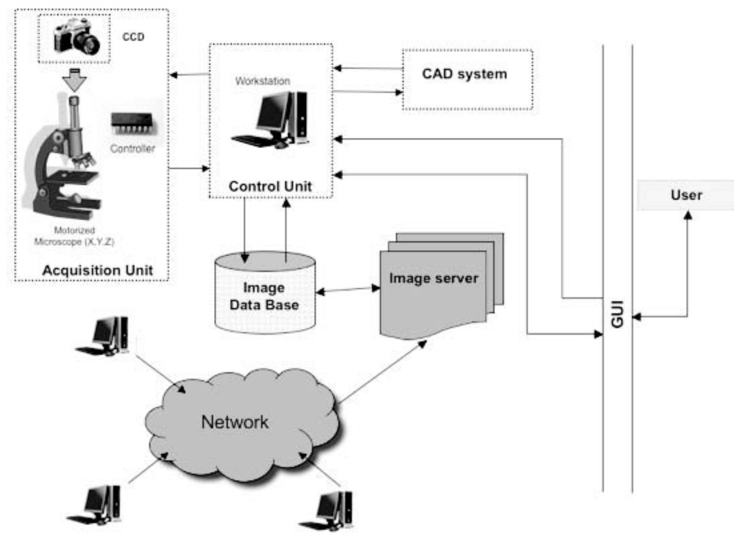
In this respect, many results have been achieved in a wide range of microscopy applications, such as the analysis of blood smears, chromosomes, tumour cells, brain, bone marrow, skin lesions, or autoantibodies analysis. These results can be attained, since the strategies in microscopy analysis have been expanded by ongoing automation of sample preparation, field selection, focusing, and imaging. Consequently, scientists and physicians are nowadays able to collect large sets of high-quality microscope images.

Figure 1 shows a typical architecture of a computer-based system devised to analyze the micrographs. The user controls and interacts with the system through a Graphical User Interface (GUI), which is located at a workstation. The workstation controls a full-motorized microscope, whose parameters, such as focusing, movement stage, objective, illumination, and magnification can be flexibly set. Moreover, the microscope is equipped with a digital camera to acquire one or more images of the sample under examination. The digital images are transferred to the workstation, and then they are automatically post-processed, stored, and shared on the Web.

Such architecture should act as a computer-aided diagnosis (CAD) system. It is well-known that such systems play an important role since they not only support the specialist in the image analysis task, but they also provide a set of useful functionalities that help in speeding up the routine part of the work.

In this article, we discuss the use of computer-based systems in microscopy, focusing on cytological images. We initially present recent results on image segmentation, and then we argue that it makes sense moving from a structural approach to a semantic interpretation of micrographs. In this respect, we focus on the relevance of using CAD tools to overcome the current limitations of microscopy, investigating several peculiar objectives of such systems. A short review of the literature demonstrates that the development of a

Figure 1. Typical architecture of a computer-aided diagnosis (CAD) system



flexible CAD applicable to various working scenarios is a future trend in microscopy healthcare systems. To support our position, we briefly describe a tool that analyzes and classifies fluorescence images.

BACKGROUND

The main contributions that healthcare information systems have given to microscope applications are basically related to the following two areas:

- **Image segmentation** (i.e., partitioning the image into regions as well as locating objects, e.g., cells)
- **Image classification** (i.e., recognizing the patterns or the objects of an image)

The classification systems may use data computed by image segmentation, thus integrating information that could be provided to the specialist to help him/her in the decision-making process.

Classically, image segmentation is defined as the partitioning of an image into not overlapping, constituent regions that are homogeneous with respect to some characteristic, such as intensity or texture. In the case of cytological micrographs, image segmentation is related to cells detection, which has to satisfy the following four main requirements:

- One micrograph has to be evaluated in a very short time to enable high throughput in screening many slides.
- Results must be reproducible to guarantee a reasonable statistical interpretation of the data.
- Robustness is needed, since cells vary in shape, density, and fluorescence properties.
- Microscope images may be taken at different magnifications so that the detection system has to be easily adaptable to cells' size changes.

Many cells segmentation methods have employed image-processing techniques dealing with domain-specific problems. Nattkemper, Ritter, and Schubert (2001) and Theis, Kohl, Kuhn, Stockmeier, and Lang (2004) use an adaptive neural classifier mapping each image point *m* to a fluorescent micrograph to a confidence value *C(m)* in [0;1], indicating how probable a cell lies in this image patch or not. This function is then used as a local filter on the whole image, providing a probability distribution with local maxima at cell positions. Analysis of maxima by thresholding reveals the number and position of the cells. Perner, Perner, and Muller (2002) and Chen, Zhou, and Wong (2006) localize fluorescent cells applying global thresholding by the Otsu's algorithm (Otsu, 1979). The algorithm well locates the cells with their cytoplasmic structure, but not the nuclear envelope itself. To overcome such limitations, Perner et al. (2002) use morphological filters, whereas Chen et al. (2006) apply watershed techniques. In both papers, overlapping cells are eliminated by a simple heuristic based on compactness evaluation. Fang, Hsu, and Lee (2003) uses local adaptive thresholding and watershed algorithm to separate the clumped cell clusters into individual objects. However, due to histological noise in the images, several false identifications are obtained that are recognized by a second stage. In (Wählby, Sintorn, Erlandsson, Borgefors, & Bengtsson, 2004) a region-based segmentation has been applied to both cervical and prostate carcinoma images. Morphological filtering combination of both the original image and its gradient magnitude creates the seeds representing both object and background that are used as starting points for watershed segmentation of the gradient magnitude image. Finally, clusters of nuclei are separated on the basis of their shape.

It is worth noting that segmentation methods based on thresholding or morphological filtering do not take into consideration the object shape. Therefore, retrieved contours tend to have irregular notches and the automatic separation of nuclei is not facilitated. In this respect, Würflinger, Stockhausen, Meyer-Ebrecht, and Böcking (2004) proposes a segmentation algorithm based on active contours. They use cubic B-Splines as an object model, whose movement is driven both by external image forces, computed from image edges, and by internal forces that model stiffness features.

The performances of segmentation algorithms are usually computed by comparing the results achieved using traditional methods (i.e., the manual one) with those obtained by automatic tools. In this respect, in the literature are reported some works that use commercial systems to detect cells (Méhes, Lörch, & Ambros, 2000; Yamamura, Yamane, Hino, Ohta, Shibata, Tsuda, & Tatsumi, 2002). In these papers, the authors report good correlation between results attained both manually and automatically. Moreover, Méhes et al. (2000) show that the determination of tumour cell content in haematopoietic sample is only reliable when it is performed together with a computerized system. The systems presented in (Chen et al., 2006; Fang et al., 2003; Nattkemper et al., 2001; Theis et al., 2004; Wählby et al., 2004; Würflinger et al., 2004) correctly detect a high rate of cells (i.e., approximately up to the 90-95% of cells). In (Nattkemper et al., 2001; Nattkemper, Twellmann, Schubert, & Ritter, 2003) for images of good quality the segmentation algorithm achieves the accuracy of the medium skilled experts, whereas in images with increased noise the computer-based method is outperformed by some of the experts.

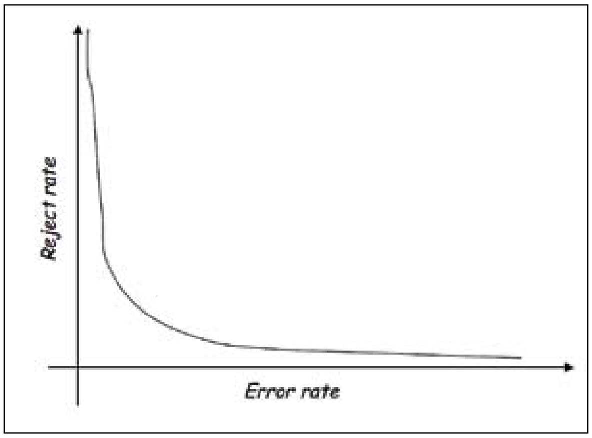
This brief literature review shows that using segmentation tools helps the specialists in the task of cells detection, providing more reliable data. Nevertheless, in several applications, physicians have to classify the images after the separation of biological objects from the background.

Up to now, during the inspection of micrographs, the physicians use only qualitative or semiquantitative information, thus limiting method reproducibility. In order to overcome these drawbacks, we deem that supporting more the specialist by moving from a structural approach to a semantic interpretation of micrographs is a promising trend. The final goal is the classification of the whole image as well as regions or objects detected by the segmentation.

To this aim, pattern recognition techniques are employed, such as those based on supervised or unsupervised approaches. Note that, in spite of several years of research and development in the field of pattern recognition, the general problem of recognizing complex patterns with arbitrary scale, orientation and location remain unsolved. Therefore these systems have to be tailored to the specific application domain as well as robust and efficient.

The ultimate performance measure of a general pattern recognition system is overall error rate (P_e). However, in many applications, such as the medical ones, the designers usually implement a *reject option* that aims at rejecting the highest possible percentage of samples that would otherwise be misclassified (i.e., misclassified without a reject option).[2] It is worth noting that, even when used adequately, this criterion introduces a side effect whereby some samples that otherwise would have been correctly classified are rejected. Invoking the reject option reduces the error rate; the larger the reject rate, the smaller the error rate. This relationship is represented as an error-reject trade-off curve that can be used to set the desired operating point of the classifier. It is monotonically nonincreasing, since rejecting more patterns either reduces the error rate or keeps it the same (Figure 2). As example, a good choice for the reject rate is based on the costs associated with reject and incorrect decisions (De Stefano, Sansone, & Vento, 2000; Golfarelli, Maio, & Maltoni, 1997).

Figure 2. Example of a theoretical error-reject trade-off curve



COMPUTER-AIDED DIAGNOSIS SYSTEMS IN MICROSCOPY

Achievements in computer vision and artificial intelligence have shown that CAD systems can be used in many field of medicine, especially in those based on digital images, such as microscopy. In general, a CAD tool provides useful data obtained by analyzing medical data by means of a computer, which can be used to support the human specialist in the diagnosis process (Cheng, Cai, Chen, Hu, & Lou, 2003; Sack, Knoechner, Warschkau, Pigla, Emmerich, & Kamprad, 2003; Takehiro, Kenichi, & Shinichi, 1992; Van Ginneken, Ter Haar Romeny, & Viergever, 2001).

Employing CAD systems, five major objectives can be pursued: (a) carrying out mass screening campaigns, (b) performing a preselection of the cases to be examined, enabling the physician to focus his/her attention only on relevant cases, (c) serving as a second reader, thus augmenting the physician capabilities and reducing errors, (d) aiding the physician while he/she carries out the diagnosis, (e) working as a tool for training and education of specialized medical personnel.

Each of those objectives has its peculiar requirements that are now further characterized and synthetically reported in Table 1.

In case (a), the CAD carries out mass screening campaigns, acting as a full-automated system that labels all input samples (i.e., as a zero-reject system), thus increasing the throughput. It is worth noting that the physician a priori knows such a rate, since the overall P_e can be experimentally evaluated.

In case (b), the physician will diagnose only cases rejected by the CAD, whereas the classified samples will not be presented to the human expert, who completely trusts in the system. Therefore, the recognition system approaches to a zero-error classifier whatever the reject rate (although in a real application it is almost impossible to not have misclassification). Indeed, with reference to a theoretical error-reject curve, the more the error approaches to zero, the greater is the reject rate (at limit 100%). Note that now the CAD approaches to a zero-error system and rejects doubtful samples. Therefore, with respect to case (a), fewer even though more accurate tests are completed.

In case (c), the CAD serves as a second reader, supplying an opinion to the physician. Now, the CAD acts as a zero-reject system, providing also a reliability measure of its decision.

In case (d), the recognition system aids the physician during the diagnosis, performing as a zero-reject system.

Finally, in case (e), the CAD acts differently on the basis of both training purpose and people skills.

It is worth noting that the recognition system can work between two extreme cases (i.e., the zero-error and the zero-reject). Therefore, the error-reject curve

Table 1. Description and requirement a CAD system for different working scenarios

Case	***Working scenario description***	***CAD requirement***
a	Mass screening campaigns	Zero-reject system
b	Samples pre-selection	Reject system
c	Second reader	Zero-reject system and reliability measure
d	Aiding expert in the diagnosis	Zero-reject system
e	Training and education of specialists	Depends on the purpose

can be used to flexibly set the system operating point between these extremes.

In the literature, several papers can be found employing computer-based tools that classify either the whole micrographs or single objects detected in them. They support the relevance of developing *intelligent* tools capable of performing semantic interpretation of micrographs. However, to our knowledge, analysis of the literature shows that CAD systems flexibly applicable to the previous different working scenarios cannot be easily found. In the following, we report some recent papers supporting such an observation.

Fang et al. (2003) integrate three classifiers (Naïve Bayes, decision-tree and association mining rule) using a majority voting strategy in order to distinguish between tumour and nontumour cells so as to improve the accuracy of the cells identification procedure (see section 2). The average P_e is 5.7%. To classify cancer cell nuclei in time-lapse microscopy, Chen et al. (2006) employs 6-Nearest-Neighbor (6-NN) classifier with a set of 7 statistical features. They apply heuristic rules on established biological knowledge of cell phases, thus correctly classifying the 98.4% of samples. Sack et al. (2003) and Perner et al. (2002) present some results on mining staining pattern of fluorescent cells in Indirect ImmunoFluorescence (IIF). Texture-based features are computed on 321 cells detected by the segmentation procedure reported in section 2. These 132 features are given to a decision tree induction program to find out the most relevant subset and to construct the classification knowledge. These systems exhibit an error rate of 25.6% (Perner et al., 2002) and 16.9% (Sack et al., 2003). The classification of cell types for the early detection of cancer has been presented in (Würflinger et al., 2004). The system distinguishes between mesothelial cells, lymphocytes, granulocytes and macrophages on the basis of eight features related to their size, form, and texture. The authors tested both Bayes classifier based on Gaussian mixture and kNN classifier. The latter performs better that the former, successfully recognizing the 87.5% of samples. Huang and Murphy (2004) presents a support vector machine system devised to classify fluorescence microscope images without segmenting them. Using a set of 26 statistical features, P_e is 5.2%. Boland (1998) demonstrates the feasibility of applying machine learning techniques to subcellular localization patterns in images of proteins and DNA obtained using fluorescence microscopy. They adopt both a backpropagation neural network and a classification tree induction. In both cases, the hit rate is 87%.

Finally, other classification system applied to microscopy can be found in Debeir, Decaestecker, Pasteels, Salmon, Kiss, and Van Ham (1999); Keller, Gader, Sohn, and Caidwell (2001); Lin, Xiao, and Micheli-Tzanakou (1998); Long, Cleveland, and Yao (2006); Milenovic, Stoiljkovic, and Stojanovic (1998); Seidenari, Pellacani, and Pepe (1998); and Weyn, van de Wouwer, Kumar-Singh, van Daele, Scheunders, van Marck, and Jacob (1999).

A deep analysis of these works shows that all of them can act only as zero-reject systems without providing a reliability measure of final classification. Therefore, with reference to Table 1, they are best suited to be applied in cases (a), (d), and (e).

In other fields of medical imaging, the introduction of a reject option allows varying the operating point of CAD (De Stefano et al., 2000; Golfarelli et al., 1997). Therefore, the range of application of CAD systems can be extended to cases (b) and (c). It is worth noting that the introduction of a reject option is usually related to the reliability measurement of each classification action since, in most cases, a sample is rejected when its reliability is smaller than a threshold (De Stefano et al., 2000).

Now, we would like to briefly discuss a recently proposed CAD tools applicable to a wide range of situations in the detection of antinuclear autoantibodies. Particularly, those CAD systems aim at classifying fluorescence intensity in IIF tests.

In this respect, both the systems proposed in Soda and Iannello (2006a) and Soda and Iannello (2006b) adopt a multiple expert system (MES) based on the classifier fusion paradigm, whereas in Soda, Iannello, and Vento (2006c) and Soda, Iannello, and Vento (2006d), the authors present two MES based on a classifier selection approach. The database used for evaluation is the same among these works and consists of 600 images; the results show that the system based on classifier selection performs better than the others.

CAD tools discussed in (Soda & Iannello, 2006a, 2006b) make use of features inspired to medical practice and show low P_e, up to 1%. Nevertheless, they use a reject option and do not cast a result in about 50% of cases. The adopted classification rule does not allow a flexible management of samples that are intrinsically hard to classify, justifying the high reject rate required to obtain low error rates. The CAD is suited to application in case (b), since it can reject or not input samples.

The MES proposed in (Soda et al., 2006c, 2006d) are made up of three specialized experts, each one devised to recognized one of the input classes. A combination module using two different rules, which provide both a fixed-reject and a zero-reject system, respectively, sets the CAD output. The first rule is based on the binary combination of the experts' outputs, whereas the second one adopts a zero-reject strategy based on reliability estimation of each classification act of the experts.

In the fixed-reject mode, the CAD exhibits an error rate of 1.33% and a reject rate of 11.33%; furthermore no samples are erroneously assigned to the negative class (0.00% of false negative). In the zero-reject mode, the hit rate is 94.33% (1.33% of false negative).

In order to understand when it is preferable to use one rule respect to the other, the authors introduce a global cost function that depends upon both the cost of a misclassification and the cost a rejection as well as the gain of a right classification.

Such a convenience analysis shows three different sets of operating points, which allow applying the CAD to all the working scenarios reported in Table 1 (Soda et al., 2006d).

FUTURE TRENDS

The field of micrograph processing involves a wide range of scientific disciplines and it is characterized by typical limitations. To overcome them, future trends in micrographs processing are directed toward both image segmentation and classification.

Hence, the current challenge is the development of CAD tools that can be flexibly set to the requirements of various working scenarios. We deem that such distinctive feature will allow using these systems in daily medical practice.

To pursue these goals it is desirable that problems across specific applications as well as terms for the evaluation of system performance will be proposed.

CONCLUSION

In this article, we discuss the use of healthcare information systems in cytological microscopy, since they have demonstrated the potential of overcoming the main limitations of microscopy. Indeed, in the last years, several results have been achieved in both image segmentation and classification, improving the medical practice.

The analysis of the literature shows that moving from a structural approach to a semantic interpretation of micrographs is beneficial. We investigated peculiar objectives of CAD tools, showing that the development of comprehensive systems that can be flexibly applied to various working scenarios is one of the current challenges in micrographs processing.

REFERENCES

Boland, M.V., Markey, M.K., & Murphy, R.F. (1998). Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. *Cytometry, 33*(3), 366–375.

Chen, X., Zhou, X., & Wong, S.T.C. (2006). Automated segmentation, classification, and tracking of cancer cell nuclei in time-lapse microscopy. *IEEE Transactions on Biomedical Engineering, 53*(4), 762–766.

Cheng, H., Cai, X., Chen, X., Hu, L., & Lou, X. (2003). Computer-aided detection and classification of microcalcifications in mammograms: A survey. *Pattern Recognition, 36*, 2967–2991.

Debeir, O., Decaestecker, C., Pasteels, J.L., Salmon, I., Kiss, R., & Van Ham, P. (1999). Computer-assisted analysis of epiluminescence microscopy images of pigmented skin lesions. *Cytometry, 37*(4), 255–266.

De Stefano, C., Sansone, C., & Vento, M. (2000). To reject or not to reject: That is the question-an answer in case of neural classifiers. *IEEE Transactions on Systems, Man and Cybernetics, Part C, 30*(1), 84–94.

Fang, B., Hsu, W., & Lee, M.L. (2003). On the accurate counting of tumor cells. *IEEE Transactions on Nanobioscience, 2*(2), 94–103.

Golfarelli, M., Maio, D., & Maltoni, D. (1997). On the error-reject trade-off in biometric verification system. *IEEE Transaction on Pattern Analysis and Machine Intelligence, 19*(7), 786–796.

Huang, K., & Murphy, R.F. (2004). Automated classification of subcellular patterns in multicell images without segmentation into single cells. *IEEE Interna-*

tional Symposium on Biomedical Imaging: Macro to Nano, 2, 1139–1142.

Keller, J.M., Gader, P.D., Sohn, S., & Caidwell, C.W. (2001). Soft counting networks for bone marrow differentials. *IEEE International Conference on Systems, Man, and Cybernetics, 5*, 3425–3428.

Lin, W., Xiao, J., & Micheli-Tzanakou, E. (1998). A computational intelligence system for cell classification. In *Proceedings IEEE International Conference on Information Technology Applications in Biomedicine* (pp. 105–109).

Long, X., Cleveland, L., & Yao, Y.L. (2006). Automatic detection of unstained viable cells in bright field images using a support vector machine with an improved training procedure. *Computers in Biology and Medicine, 36*(4), 339–362.

Méhes, G., Lörch, T., & Ambros, P.F. (2000). Quantitative analysis of disseminated tumor cells in the bone marrow by automated fluorescence image analysis. *Communications in Clinical Cytometry, 42*(6), 357–362.

Milenovic, D., Stoiljkovic, L., & Stojanovic, N. (1998). Cell classification for diagnostic of reactive histocytichyperplasia using neural networks. In *Proceedings of 9th Mediterranean Electrotechnical Conference* (Vol. 2, pp. 1466–1470).

Nattkemper, T.W., Ritter, H., & Schubert, W. (2001). A neural classificator enabling high-throughput topological analysis of lymphocytes in tissue sections. *IEEE Transactions on Information Technology in Biomedicine, 5*(2), 138–149.

Nattkemper, T.W., Twellmann, T., Schubert, W., & Ritter, H. (2003). Human vs. machine: Evaluation of fluorescence micrographs. *Computers in Biology and Medicine, 33*(1), 31–43.

Otsu, N. (1979). A threshold selection method from gray-level histograms. IEEE Transactions on *Systems, Man, and Cybernetics, 9*(1), 62–66.

Perner, P., Perner, H., & Muller, B. (2002). Mining knowledge for HEp-2 cell image classification. *Journal Artificial Intelligence in Medicine, 26*, 161–173.

Sack, U., Knoechner, S., Warschkau, H., Pigla, U., Emmerich, F., & Kamprad, M. (2003). Computer-assisted classification of hep-2 immunofluorescence patterns in autoimmune diagnostics. *Autoimmunity Reviews, 2*, 298–304.

Seidenari, S., Pellacani, G., & Pepe, P. (1998). Digital videomicroscopy improves diagnostic accuracy for melanoma. *Journal of the American Academy of Dermatology, 39*, 175–181.

Soda, P., & Iannello, G. (2006a). A multi-expert system to classify fluorescent intensity in antinuclear autoantibodies testing. *Computer Based Medical Systems, IEEE Computer Society*, 219–224.

Soda, P., & Iannello, G. (2006b). Experiences in ANN-based classification of immunofluorescence images. *Enformatika Transactions on Engineering, Computing and Technology, 14*, 252–257.

Soda, P., Iannello, G., & Vento, M. (2006c). Using multiple experts systems to classify fluorescence intensity in antinuclear autoantibodies analysis. *Seconda Conferenza Italiana sui Sistemi Intelligenti (CISI).*

Soda, P., Iannello, G., & Vento, M., (2006d). A multiple experts system for classifying fluorescence intensity in antinuclear autoantibodies analysis. *Pattern Analysis & Applications,* accepted for publication.

Takehiro, E., Kenichi, K., & Shinichi, Y. (1992). Computer-aided diagnosis system for medical use. Patent Number: EP0487110. Retrieved February 14, 2008, from http://ep.espacenet.com

Theis, F.J., Kohl, Z., Kuhn, H.G., Stockmeier, H.G, & Lang, E.W. (2004). Automated counting of labelled cells in rodent brain section images. In *Proceedings of BioMED*. Innusburck, Austria.

Van Ginneken, B., Ter Haar Romeny, B., & Viergever, M. (2001). Computer-aided diagnosis in chest radiography: a survey. *IEEE Transactions on Medical Imaging, 20*, 1228–1241.

Wählby, C., Sintorn, I.M., Erlandsson, F., Borgefors, G., & Bengtsson, E. (2004). Combining intensity, edge and shape information for 2D and 3D segmentation of cell nuclei in tissue sections. *Journal of Microscopy, 215*(1), 67–76.

Weyn, B., van de Wouwer, G., Kumar-Singh, S., van Daele, A., Scheunders, P., van Marck, E., & Jacob, W. (1999). Computer-assisted differential diagnosis of malignant mesothelioma based on syntactic structure analysis. *Cytometry, 35*(1), 23–29.

Würflinger, T., Stockhausen, J., Meyer-Ebrecht, D., & Böcking, A. (2004). Robust automatic coregistration, segmentation, and classification of cell nuclei in multimodal cytopathological microscopic images. *Computerized Medical Imaging and Graphics, 28*(2) 87–98.

Yamamura, R., Yamane, T., Hino, M., Ohta, K., Shibata, H., Tsuda, I., & Tatsumi, N. (2002). Possible automatic cell classification of bone marrow aspirate using the CELL-DYN 4000® automatic blood cell analyzer. *Journal of Clinical Laboratory Analysis, 16*(2), 86–90.

KEY TERMS

Computer-Aided Diagnosis (CAD) System: Computer-based system that analyzes medical data in order to support the human specialist in the diagnosis process.

Features: A set of attributes that represents a pattern.

Image Classification: Recognition the patterns or the objects of an image.

Image Segmentation: Partition of an image into not overlapping, constituent regions that are homogeneous with respect to some characteristic. Ideally, a segmentation method finds sets of points that correspond to distinct structures or regions of interest in the image.

Micrographs: Microscope images.

Multi-Expert System (MES): Aggregation of classification systems according to a combination approach, in order to improve the recognition performance.

Reject Option: Classification option that aims at rejecting the highest possible number of samples that would otherwise be misclassified (i.e., misclassified without a reject option). It is usually based on the estimation of the classification reliability.

ENDNOTES

1. In the following, microscope images are referred to as *micrographs*.
2. In the following, a classification system that makes a decision on all input samples will be referred to as *zero-reject*, as *reject system* otherwise.

Restructuring a Military Medical Department Research Center

W.F. Lawless
Paine College, USA

Joseph Wood, LTC
Fort Gordon, USA

Hui-Lien Tung
Paine College, USA

INTRODUCTION

This case study is of a military medical department research center (MDRC) with access to advanced information systems (IS), yet struggling to determine the quality of its residents in training and their scholarly productivity (see the article on "Theory Driven Organizational Metrics" in this encyclopedia). Based on theory, this case study was guided by stories captured from MDRC in the collapse of four interdependent variables: planning, execution, resources, and time. Our primary goals for this case study were to: (1) Formulate recommendations to utilize the IS available to reduce the overall operational cost of MDRC; (2) increase the operational efficiency and growth of MDRC by enhancing its ability to attract new extramural funds; and (3) further explore the link between practice and theory. To the extent possible, all organizational names, references, and locations have been revised to fictitious ones.

Organizational Background[1]

The Medical Department Research Center (MDRC) is a regional military medical hospital located stateside in the USA. MDRC supports clinical and basic research for its staff and all hospital practitioners including family medicine, internal medicine, general and orthopaedic surgery, and dentistry (endodontics, peritonitis, oral maxillofacial, orthodontics, and nurse anaesthesiology). In addition to providing basic research support for the hospital, MDRC is responsible to teach the fundamentals of experimental research to the hospital's medical residents (residents working within a specialty, residents rotating among the hospital's different specialties, and interns in transition; in JWO, 2005, pp. 4), to provide continuing education for more experienced care providers, and to train dentists in research methods. One of the goals of the training provided by MDRC is to help the hospital's Graduate Medical Education (GME) candidates become certified by their respective American specialty boards. GME trainees are supervised by about 150 teaching staff members who are all board-certified in their respective specialties (JWO, 2005,p p. 4). MDRC's published goal is to provide the "best support possible for all research endeavours[…]by providing the facilities[…]and atmosphere of inquiry necessary to support and stimulate both basic and clinical medical investigations" (JWO, 2005, pp. 5).

The staffing level at MDRC is provided in Table 1. It is headed by a Chief and Assistant Chief, both military officers, with the staff separated into military (officers and enlisted) and civilian positions (full-time, part-time, and contractors) along functional lines (protocol coordination, veterinarians), and minimally by levels of education (Ph.D.).

Included among the responsibilities of MDRC are the certification and regulation of responsible practices in human and animal research. MDRC discharges its responsibility with mentors teaching residents about research, monthly Institutional Review Board (IRB) meetings for the approval of human clinical studies, and monthly meetings of the Institutional Animal and Care Use Committee (IACUC) for the approval of investigations with animals. Its IRB program has maintained "Federal Wide Assurance" certification that its training programs meet the strict federal guidelines approved by the National Institutes of Health (NIH). Its veterinary facility has maintained certification from the Association for Assessment of Laboratory Animal

Table 1. Staff

	2004[2]	2005[3]
MDRC Laboratory		
Full-Time Laboratory Staff, Stated[4]	15	15
Full-Time Laboratory Staff, Counted	19	16
Protocol Coordinators	2	2
N, Civilian Ph.D.'s	4	4
N, Military Ph.D.'s	1	1
N, Veterinarians	2	2
Specialized Contractors, Counted	3	1
Total *N* by Staff, Counted	22	17
Laboratory Staff by Rank		
Officers	5	4
Enlisted	5	4
Civilians	9	8
Contract	3	1
Total *N* by Rank, Stated & Counted	22	20

Table 2. Residents and protocols, stated

	2004[5]	2005[6]
Full-Time Residents Trained	74	72
Full-Time Residents and Fellows with approved protocols	25	14
Full-Time Resident Disciplines, listed	6[7]	5
NonMC officer Trainees with approved protocols, $N_{Trainees}$	13	11
Approved protocols for NonMC Trainees, $N_{Protocols}$	13	10
NonMC Trainee Disciplines	3	4
Hospital Staff with approved protocols, N_{Staff}	98	63
Approved protocols for Hospital Staff, $N_{Protocols}$	56	41

Care (AALAC, a nonfederal advisory group) that its animal investigations also meet strict federal guidelines (viz., AALAC conducts peer reviews, but it has no direct regulatory role). AALAC approval indicates a superior program, and that the institution meets all the rules and regulations of the USDA, Public Health Service (PHS), and other regulatory agencies. IRB and IACUC meet monthly to review research proposed in Protocols submitted by official documents along with email copies (DWA, 2004; JWO, 2005). Based on these meetings some of the trainee Protocols are rejected or are required to be revised and resubmitted (INC, 2006, pp. 10).

The Annual Reports state that GME has about 75 physicians and dentists in training in 2004 and 2005 (DWA, 2004, pp. 4; JWO, 2005, pp. 4); however, the actual counts provided in these reports were 74 and 72.

In 2005, MDRC continued to expand its support for GME scholarly activities and its search for extramural funds (JWO, 2005, pp. 4). MDRC conducts three types of research at its facility: basic, clinical, and case studies (INC, 2006). MDRC assists resident trainees by providing research assistants to help them with designing, executing, and tracking their research protocols. However, the current IS is not used in this process, nor is the process automated at any point.

The discretionary funds provided for research average about $140,000 per year (INC, 2006). Extramural grants are an additional source of funds, currently averaging $3–400,000 per year. In 2005, MDRC received $800,000 in extramural funding for laboratory-based research (JWO, 2005, pp. 4).[8]

Some of the research supported by MDRC is performed voluntarily, but a larger fraction is required to be performed: "[…]supporting clinical and basic research for all [hospital] personnel [is] now a requirement for matriculation in many training programs" (DWA, 2004, pp. 4; JWO, 2005, pp. 4). And, "Student nurse anesthetists are required to complete a research project for publication" (DWA, 2004, pp. 4). However, the impact of making research a requirement on the scholarly activities of trainees is unknown.

CASE DESCRIPTION[11]

In 2006, a new director assumed command of MDRC (INC, 2006). Over the remaining two years of his administration, the Chief of MDRC wants to establish MOE's (Metrics of Effectiveness) to measure his organization's success and to craft a plan with MOEs to improve the performance of scholarly activities (i.e., plan, strategy, or Business Model, (BM)). Some resident trainees start their research rotation prepared with a line of investigation derived from their own interests, collaboration with peers, or previous mentors. But if they are not presently working in research or have a research interest, a mentor is assigned to them by MDRC staff. One problem with using MOEs is that much of the research proposed in the MDRC protocols has lasted or may last for a number of years before scholarly products are published, whereas other research protocols may last much less than one year, giving the less complex Protocols an advantage in the generation of scholarly products (see Table 4). As the complexity of a Protocol increases, the time necessary

Table 3. Funding, in thousands of dollars

	2004[9]	2005[10]
Military, Total	$319.3	$319.3
Civilian	507.1	544.4
Consumables	65.2	85.2
Civilian contractors, consultants	32.7	42.4
CEEP	0.0	269.1
TDY Staff	9.7	42.2
Paper Presentations	28.3	1.3
Publications	0.0	0.0
Civilian, Total	643.0	984.6
Grants:		
Extramural Funds, Listed		
G: 273.2	273.2	83.2
H: 58.8	58.8	0.1
T: 63.6	63.6	91.0
Extramural Funds, Secured, Stated	0.0	800.0
Grants, Total	395.6	319.3
Total	$1,357.8	$1,453.0

Table 4. Protocol dispositions using an ad hoc impact score

	2004[12]	2005[13]
Published Manuscripts		
Protocols-to-Publications	172/87 = 1.98:1	156/74 = 2.11:1
Index (PPI)	284	157
Publication Rank[14] [15]	-193	-59
Author Position[16]	91/34=2.68	98/18=5.44
Average Score		
Presentations		
Presentation Rank[17]	122	102
Author Rank[18]	-66	-9
Average Score	56/36=1.56	93/45=2.07
Abstracts		
Abstract Rank[19]	17	15
Author Rank[20]	-54	-44
Average Score	-37/17=-2.18	-29/11=-2.67
Overall Average Score	2.06	4.8

to complete a program of research also increases. The Chief wants to increase the complexity of protocols, but remain able to measure the impact of complexity on scholarly productivity.

The Chief of MDRC wants to hire additional personnel, but justifications for additional personnel should be based on MOEs and not intuition. MDRC's overall status with all of its funding resources is also unknown. But without clear MOEs, the decision by the administrators has been that potentially 10–20% of new research is turned away. The Chief believes that better MOEs could justify more personnel to administer additional internal and extramural grants.

The Chief is required to give command briefings on productivity (INC, 2006). One administration goal is to make better presentations for Command briefings. In these presentations, the Chief would like to be able to compare his organization's performance with other MDRCs in his region and nationally. In Table 5, data are presented on the performance of MDRCs from other sites and regions across the nation. (per protocol, for the productivity based on the number of scholarly activities, number of personnel, facility space, and total funding).

CURRENT PROBLEMS

Productivity

Sometimes, the ratio of the number of protocols to publications can get as high as about 5:1. The Chief wants MOEs to better justify budget and staff at the beginning of the year, and to determine and track scholarly and funding trends during the year. He believes that scholarly activities have increased while funding has held more or less constant. However, one trade off possible is that the lack of focus among trainees precludes more protocols (INC, 2006, pp. 3–4,7). But he is hesitant to change his business model. If he pushes for increased complexity, it would likely result in even fewer manuscripts being published. Education is also a goal, but a MOE for education is problematic; however, the impact of scholarly products may help to indirectly determine the level of education and training of trainees. Scholarly productivity currently treats abstracts, manuscripts, and presentations equally. The picture presented in Table 5 could change dramatically once journal impacts are incorporated (e.g., see Garfield, 2006).

Technology Components

MDRC is located stateside at one of the American Department of Defense's (DoD) major military installations. Its primary mission is to ensure force readiness and mobilization in order to project power globally from its platform in the U.S. part of its responsibilities means that this military base serves as a major source of IT R&D for DoD. This also means that telemedicine and distance learning, directed by the base, have become a global military resource for all U.S. armed forces. The Chief implied that upon his request, the hospital's Division of Management Information will provide

Table 5. Between-group (clinic) comparisons of seven military medical research centers

Ranked By Group Size	FY2004[21]	FY2005[22]	FY2005 Productivity Impact Index
Unit 1	2.37; 0.084; 39.96; 35,249	2.44; 0.08; 40.04; 14,961	9.44
Unit 2	1.07; 0.087; 185.8; 14,622	1.71; 0.085; 200; 15,545	5.12
Unit 3	1.06; 0.086; 91.7; 24,991	1.12; 0.103; 98.45; 21,727	5.57
Unit 4	1.59; 0.128; 127.6; 12,478	1.23; 0.116; 112.5; 19,070	5.06
MDRC	1.13; 0.314; 112.8; 18,241	.85; 0.214; 115.5; 12,915	3.7
Unit 6	2.74; 0.645; 523.4; 54,363	3.27; 0.73; 540.9; 52,900	5.27
Unit 7	0.86; 0.086; 0.0; 59,003	0; 0.057; 0; 10,506	3.97
Overall Group Average	1.60; 0.121; 109.3; 28,244	1.52; 0.198; 184.6; 21,089	NA

limited support to implement the recommendations made in this study.

Organizational Concerns

As depicted in Table 5, there is wide variability in the time to complete a protocol. Protocols may take 3–10 weeks to complete, or 3–10 years (INC, 2006). In addition, there are no MOEs to distinguish their intrinsic value or performance (effectiveness and efficiency). The Chief and his Assistant would like to craft a MOE to determine performance based on data. For example, there are fewer protocols in 2006, but whether it makes a difference to their organization's performance or the research it supports are both unknown.

60% of MRDC's research support is derived from extramural grants (INC, 2006). But there are no guidelines to conduct a search for extramural funding or to measure its impact on MDRC's performance, increasing management uncertainty as a result.

Extramural finding appears not to be calibrated into mission plans or performance objectives (INC, 2006). Most of the funds MDRC relies on to perform research are provided in the military budget process, but it has remained relatively fixed. The Chief likens his Division to a research university in that it performs important research for the military and education for hospital GME trainees, but he believes that the process should be more focused.

Management Issues

MDRC is in a state of flux (INC, 2006). Overall, it is difficult to restructure (changing its Business Model), to distribute research funds equitably, or to determine what to do based on the popular literature (e.g., Allen, 2002). Since theory for organizations (Weick & Quinn, 1999) has failed, not surprisingly, data for restructuring outcomes is poor (Kohli & Hoadley, 2006). The end result is an ad hoc or shotgun approach of BPR. In early 2006, the two chief administrators began to consider how to improve the process. From Table 4, some protocols written 2–3 years ago have shown little progress. Presently, there are about 130 open protocols in 2006. But problems exist with MDRC's published data that raise questions about its reliability; for example, not all abstracts are in the FY05 data, meaning that the Annual Report understates performance. The Chief estimates that there are about 60 abstracts per year, but only about 10 were captured by his division in 2005 data (JWO, 2005). The database is currently unable to distinguish between abstracts and posters; further, the number of presentations made by trainees and staff are unknown, especially to outsider groups and organizations (INC, 2006). A correct accounting would help IRB as well as the military's public affairs to get their story out to advertise better. His division has been directed to increase the number of Abstracts published by hospital residents. But some manuscripts are written only after three years, due to complex protocols, plus he has about 25% fewer staff than last year. He is also unable to discriminate between case-study driven Protocols and laboratory bench studies. Protocols are generated by standard forms of 5–30 pages (but all information is either not collected or published; see a sample Protocol Data Sheets in DWA, 2004; JWO, 2005).

The Chief is uncertain about how the mentors external to MDRC are chosen (INC, 2006), their quality and their expertise; for example, one mentor can finish a given protocol in one month, while another takes about eight months to complete a similar protocol. Productivity of the hospital mentors is currently not measured.

Residents are assigned to MDRC for either six months full-time duty (INC, 2006), or one year at about 70% of their time; others are assigned to research duty for two weeks at MDRC before returning to two weeks of duty at the hospital, but the impact of these different assignments is currently not measured. The total number of residents at the hospital is unknown. All residents are placed on a Protocol, but their performance is a function of how rapidly they generate data in either the laboratory or with a case study.

Currently, MRDC is meeting all of the needs of hospital residents, but how well is unknown (INC, 2006). However, setbacks occur during a protocol that impede the execution of research in the short-term, and have unknown effects on the performance of research over the long-term. Some protocols have been rejected to minimize administrative costs and laboratory setbacks, requiring revisions by hand, but this process is not tracked. The internal seminars sponsored by other divisions are not tracked; however, mentors are supposed to review research before a presentation, even before an internal seminar is given. Each year, one internal seminar is sponsored and judged by MRDC, leading to awards given for the best research. Generally, it takes about one year to educate a resident on the research process.

Some residents are required to perform research at MDRC as part of their rotational duties, but a few residents actually want to participate in research projects (INC, 2006).[23] Most residents do case studies; only one resident in the last six years did laboratory research, whereas almost all dental residents perform laboratory bench studies. However, the research performed by the dental residents is usually setup with minor variations among their Protocols, even though dentists push for publications. The Chief's office supports dentists, but has little control over their Protocols.

In the Chief's opinion, attracting better quality residents will require an increase in the quality of publications and getting help from other Division Chiefs, plus having a stronger program (INC, 2006). Better graduates will attract other quality candidates, adding credibility to the BM.

Funding Concerns

Lots of clinical trials sponsored by industry occur on base, implying more research funds for the hospital and MDRC (INC, 2006). Industry grants are part of MDRC's Protocol data (he estimates about 50-60 of 130 of the ongoing Protocols), but MDRC is not tracking the manuscripts generated from these extramural grants. Industry grants often operate with 3–5 year clinic trials at multiple research centers by large research teams. Clinical trials tap into the local population, but data on local pathology is often unavailable and unknown; for example, the number of local diabetics is unknown (INC, 2006). Marketing to industry is good for the hospital. The MRDC Chief is a member of the local city's Research Alliance (including local state and other hospitals) working to attract industry to the area.[24] Seed money provided by the local Medial Association indicates an interest in determining the medical assets available in the area, the size of the local population, the scholarly output in the local area, and operational budgets. However, industry is not tracked in the database.

Equipment and Facility Concerns

Some equipment costs from $20,000–40,000 each requires maintenance contracts, yet sits in the laboratory unused, producing an inefficient use of funds and space (INC, 2006). Funds for medical and nonmedical technology replacement come from the Capital Expense Equipment Program fund (CEEP). The Chief has staff in additional offices at the hospital; the main facility for MRDC is located about one-half mile away from the hospital.

RECOMMENDATION

Throughout the case study, 30 recommendations using IT were reviewed and presented to MDRC management to improve its operational performance (for a full listing, see SSRN website). In general, these recommendations were designed to automate business processes, to reduce waste, and to minimize the loss of information pertinent to strategic decision-making. Every research protocol needs to be drafted, revised and resubmitted for approval then critiqued by the IRB/AALAC as necessary. Automating this process with an Electronic

Protocol System (e-PS) will save time and capture data that can help to identify and remove potential barriers to success (e.g., recently, speedup occurred subsequent to this review by making all submissions electronic, and making research starts not wait for post-IRB/AAUC command re-review). PMS will track due dates, research progress and project completion. The next step will be an automatic document management system (DMS). DMS will track document creation (manuscripts), revision, and submission status among organization internal reviewers and external publication reviewers. In addition, DMS will track author credit and journal impacts as well as providing a document storage-retrieval feature. The more important aspects of ePS and DMS will be the capability to compile data for administration functions (e.g., annual reports, organizational reports, audits), and for real-time online metrics.[25] The metrics will be designed to measure organizational by-in for each military department and support among MDRC members and trainees for the new IT process; how well the MDRC executes its plan as an organization; whether new resources are being brought into MDRC or being saved by MDRC; and how quickly MDRC is executing on average and responding to new opportunities.

SUMMARY

Fragmentation from a lack of focus among the members of an organization indicates a general reduction in motivation to serve the organization, and consequently the need to change a business model to refocus an organization's strategy. The present case study of MDRC indicated that fragmentation among its processes and researchers had reduced control over its future plans. On the one hand, fragmentation of an organization is associated with increased innovation at the individual level (e.g., Benz, 2006), but often at the expense of the organization. Successful organizations are entangled by Command Decision-Making (CDM) into centers of cooperation (Lawless, Bergman, & Feltovich, 2006a; Lawless, Bergman, Louca, & Kriegel, 2006b) to acquire new resources and reduce wasteful expenditures (Smith & Tushman, 2005), but that also tends to reduce innovation from their marginalization of new knowledge in making the organization; in this case, MDRC, more competitive in its marketplace for new grants, improved mentors, better trainee candidates, and new joint ventures with industry or other research facilities. But this reduced innovativeness can be countered by mindfully taking steps with IT to free MDRC from many labor-intensive activities by automating data collection and analytical processes in order to free resources to increase innovation. The overall goal of implementing the recommendations is to reduce the overall costs of MDRC, increase its operational efficiency, to increase its growth by enhancing its ability to attract new extramural fund, and to reduce wasted effort in order to improve organizational effectiveness as a means of saving scarce resources.

REFERENCES

Allen, D. (2002). *Getting things done: The art of stress-free productivity*. Reprint. Penguin.

Benz, W. (2006). *A concise history of the Third Reich*. Dunlap, T. (Trans.). UCLA Press: Berkeley.

Garfield, E. (2006). The history and meaning of the journal impact factor. *JAMA, 295*, 90–93.

Hardin, G. (1968). The tragedy of the commons. *Science*, *162*, 1243-1248.

INC. (2006). Interview notes dated June 2, 2006, and subsequent informal communications.

Kohli, R., & Hoadley, E. (2006). Towards developing a framework for measuring organizational impact of IT-enabled BPR: Case studies of three firms. *Data Base for Advances in Information Systems, 37*(1), 40–58.

Kruglanski, A. W., Pierro, A., Mannetti, L., & De Grada, E. (2006). Groups as epistemic providers: Need for closure and the unfolding of group-centrism. *Psychological Review*, *113*, 84-100.

Lawless, W.F., Bergman, M., & Feltovich, N. (2006). A physics of organizational uncertainty. Perturbations, measurement and computational agents. In Shu-Heng Chen, Lakhmi Jain, & Chung-Ching Tai (Eds.), *Computational Economics* (pp. 268–289). Hershey, PA: Idea Group Inc.

Lawless, W.F., Bergman, M., Louçã, J., Kriegel, N.N. & Feltovich, N. (2007). A quantum metric of organizational performance: Terrorism and counterterrorism. *Computational & Mathematical Organizational Theory*, *13*, 241-281.

Lawless, W. F., & Grayson, J.M. (2004). A quantum perturbation model (QPM) of knowledge and organizational mergers. In L. van Elst, & V. Dignum (Eds.), *Agent mediated knowledge management*. Berlin, Springer (pp. 143-161).

Nowak, M. A., & Sigmund, K. (2004). Evolutionary dynamics of biological games. *Science*, 303, 793-799.

Smith, W.K., & Tushman, M.L. (2005). Managing strategic contradictions: A top management model for managing innovation streams. *Organizational Science, 16*(5), 522–536.

Weick, K.E., & Quinn, R.E. (1999). Organizational change. *Annual Review of Psychology, 50*, 361–386.

Wood, J. (2004). *MDRC Clinical Investigation Program Annual Report 2004: RCS MED-300 (R)*. For access, contact joseph.c.wood@us.army.mil

Wood, J. (2005). *MDRC Clinical Investigation Program Annual Report 2005: RCS MED-300 (R)*. For access, contact joseph.c.wood@us.army.mil

KEY TERMS

Agent-Based Model (ABM): Used to create complex models of organizations and systems; for example, with Monte Carlo methods, ABM's can be designed to reduce biases in Critical Path Method (CPM models and Program Evaluation Review Technique [PERT]) models.

Business Models (BM): A business model is a complex plan for business operations and strategy that, when it converges into a consensus, it maximizes productivity, but the stronger the consensus in support of the BM, the more that innovation is reduced. Alternatively, divergence increases opportunities for innovation, but by reducing productivity. This paradox is reduced by creating ambidextrous organizations that can operate in a state of tension between both increasing productivity and innovativeness (Smith & Tushman, 2005).

Command Decision Making (CDM): Top-down decision-making, especially the autocratic decision-making practiced in industry, business, and the military, but also by dictators. CDM reduces innovativeness but increases productivity. Further, CDM often employs consensus rules in its decision-making processes, because CR is open to exploitation (Kruglanski et al., 2006).

Consensus-Seeking Rules (CR): By reducing evidentiary barriers to discussion, consensus-seeking rules often lead to the least common worldview amongst risk perceivers, making it unlikely to reach a practical decision other than to instantiate past organizational practices, but consequently, reducing innovativeness. Further, consensus rules require considerable time to listen to each risk perception.

Majority Rules (MR): Majority rules often lead to faster decisions, more practical decisions, and, counterintuitively, stronger consensuses. The problem with majority rules is that they introduce conflict into decision-making. But if the conflict can be moderated or managed, the result is more learning among the participants compared to consensus rules.

Measurement Problem/Paradox: Assuming the existence of uncertainty in the four interdependent variables of planning-execution and resources-time, decreasing the uncertainty in either set raises the uncertainty in its correspondingly linked interdependent variable.

Methodological Individualism (MI): Game theory (Nowak & Sigmund, 2004) has attempted to substantiate the superiority of cooperation, based on the belief that cooperation leads to the highest social good, even if it is coerced (Hardin, 1968). However, the lack of substantiating evidence in support of the social worth of cooperation is reviewed in Lawless and Grayson (2004).

Metrics: The metrics for MDRC are being designed to convert the measurement problem into an organizational metric of performance.

ENDNOTES

1. Taken from Annual Reports; e.g. JWO, 2006; JWO, 2005.
2. 2004 Annual Report (DWA, 2004).
3. 2005 Annual Report (JWO, 2005).
4. Stated but not further identified in the Annual Reports. The business manager is part time but not so noted in the Annual Reports

5 2004 Annual Report (DWA, 2004).

6 2005 Annual Report (JWO, 2005).

7 Includes Child Psychiatry.

8 These additional grant funds were not further discussed in the Annual Report.

9 2004 Annual Report (DWA, 2004).

10 2005 Annual Report (JWO, 2005).

11 Based on interviews with the Chief and Assistant Chief administrators.

12 2004 Annual Report (DWA, 2004).

13 2005 Annual Report (JWO, 2005).

14 Journal publication ranks based on impacts have not been used in this study due to a lack of publication information, but they are available once publication information is collected (Garfield, 2006). Also, see Journal Citation Reports (JCR): portal.isiknowledge.com/.

15 Journal Publications score, Counted: +10 if Major, +3 if State, +1 if on Site, +5 if Major but no Volume listed.

16 Multiple Journal Authors score, Counted: -1 per Multiple Author, +2 for Single Authors in Major Journals, +1 for Single Authors in Minor Journals.

17 Presentations score, Counted: +5 if National, +2 if State or Regional, +1 if Site.

18 Multiple Presentation Author score, Counted: -1 per Multiple Author, +2 for Single Author Annual Meeting, +1 for Single Author State/Regional Meeting, +0 for Single Author Onsite Meeting.

19 Abstracts score, Counted: +3 if Major, -2 if only Submitted, +2 if State or Database, +1 if Site or Not Identified.

20 Abstracts authors score, Counted: -1 per Multiple Author, +1 for Single Authors, +0 if publication Not Identified.

21 Based on DoD data, 2004.

22 Based on DoD data, 2005.

23 The different resident categories should be captured in the database to contrast the productivity created by these differences.

24 See theory article for a discussion about the importance of resources and innovation (Lawless, Wood, & Tung, 2007, this volume).

25 See theory article.

R

RFID as the Critical Factor for Superior Healthcare Delivery

A. Dwivedi
University of Hull, UK

T. Butcher
University of Hull Logistics Institute (UHLI), UK

INTRODUCTION

Innovations in information and communication technologies (ICTs) have transformed the manner in which healthcare organizations function. Applications of concepts such as data warehousing and data mining have exponentially increased the amount of information that a healthcare organization has access to. Work flow and associated Internet technologies are being seen as instruments to cut administrative expenses. Specifically designed ICT implementations, such as work flow tools, are being used to automate the electronic paper flow in a managed care operation, thereby cutting administrative expenses (Dwivedi, Bali, & Naguib, 2005, p. 44; Latamore, 1999).

These recent innovations in the use of ICT applications in a healthcare context have altered the manner in which healthcare institutions exploit clinical and nonclinical data. The pendulum has shifted from the early 1980s, wherein the emphasis of ICT solutions for healthcare was on storage of data in an electronic medium, the prime objective of which was to allow exploitation of this data at a later point in time. As such, most of the early 1980s ICT applications in healthcare were built to provide support for retrospective information retrieval needs and, in some cases, to analyze the decisions undertaken. Clinical data that was traditionally used in a supportive capacity for historical purposes has today become an opportunity that allows healthcare stakeholders to tackle problems before they arise (Dwivedi et al., 2005).

However, simultaneously, a number of studies have noted that most information in healthcare is stored in silos, which do not interact efficiently with each other. Kennedy (1995, p. 85) has quoted Keever (a healthcare management executive) who notes that "Healthcare is the most disjointed industry … in terms of information exchange.... Every hospital, doctor, insurer and independent lab has its own set of information, and … no one does a very good job of sharing it."

This problem is being further acerbated by the fact that healthcare managers are being forced to examine costs associated with healthcare and are under increasing pressure to discover approaches that would help carry out activities better, faster, and cheaper (Davis & Klein, 2000; Dwivedi, Bali, James, & Naguib, 2001; Dwivedi et al., 2005; Latamore, 1999). Consequently, the expectations from modern IT applications in healthcare are for applications which support the transfer of information with context. This, in turn, has led to the emergence of clinical information systems that are led by mobile computing technologies (Dwivedi, Bali, & Naguib, 2007; Dwivedi, Wickramasinghe, Bali, Naguib, & Goldberg, 2007; Meletis, Dwivedi, Gritzalis, Bali, & Naguib, 2006).

BACKGROUND

The last decade has seen the rapid emergence and acceptance of healthcare information systems that support the concept of telemedicine and use technologies like Personal Digital Assistant (PDA), Radio Frequency IDentification (RFID) and other mobile computing technologies.

This trend has also been supported by a longitudinal survey (see Table 1) of over 200 U.S. healthcare organisations carried over a three year period, from 2000 to 2002 (Morrissey, 2000, 2001, 2002). As seen in Table 1, cClinical information systems in conjunction with mobile computing have become priority areas for healthcare institutions (see Table 1).

Modern day IT applications in healthcare, centred on mobile computing devices like PDA, RFID, and wireless local area network (WLAN) products, have already demonstrated their potential and financial vi-

Table 1. Adapted from Modern Healthcare's annual survey of information system trends in the healthcare industry (Dwivedi, Wickramasinghe et al., 2007; Meletis et al., 2006; Morrissey, 2000, 2001, 2002)

Year	2000	2001	2002
Number of healthcare organizations surveyed	224 healthcare organizations	212 healthcare organizations	255 healthcare organizations
Clinical Use of Web technology (Intranets)	60% - felt that could IT could facilitate data exchange among caregivers, that is, physician ordering of tests and access to test results	Low interest in maintaining a patient's personal health record accessible via the WWW and matching patients with clinical research. However there is renewed importance of addressing changes in this area due to regulatory obligations	Despite acknowledging that medication interaction and dosing alerts are possible within most IS - implementation has not commenced The few organizations who had made big investments in different HIS (EPR and pharmacy) are reporting substantial returns
General Uses of Web and Intranet technology	Limited use as shown by the following 15% - to share clinical guidelines 13% - to access multiple databases simultaneously 33% - as a bridge to other information systems 40% - for network wide communication of any kind	Some early success from linking "billing and insurance-query operations to payers via the Web" "Significant interest …in using the Web to improve data exchange with physicians and their office staff" About 50% indicated that they had no plans to try anything Web-related in the care-management area	33% - Using existing clinical and financial information sources to construct data repositories so as to that help spot trends and improve decision-making Further 22% are working to implement such practices whilst about 13% plan to start implementation of similar activities within a year

abilities in a healthcare context (Dwivedi, Bali et al., 2007; Dwivedi, Wichramasinghe et al., 2007; Meletis et al., 2006). Recent studies by Meletis et al. (2006) have noted that WLAN-based mobile computing allows healthcare workers to interact in real-time with the hospital's host computer system to enter, update, and access patient data and associated treatments from all clinical departments (Meletis et al., 2006).

A survey of WLAN healthcare installations found that 97% of customers indicated that "WLANs met or exceeded their expectation to provide…a competitive advantage" and that "if the productivity benefits are measured as a percentage return on the total investment … the return works out to be 48%" (McCormick, 1999, p.13). The use of PDAs by physicians has witnessed rapid acceptance in recent times. Today, about 40% of all physicians use PDAs (Serb, 2002). However, the majority of physicians are using PDAs to perform static functions. Most of them use PDAs to collect reference material with:

> *the most popular method being ePocrates - a drug reference application physicians can look up drugs by name or diagnoses, cross-reference similar medications or generic alternatives, and receive alerts on interactions…and which the Journal of the American Medical Association has described as indispensable.* (Serb, 2002, p. 44)

A few pioneering physicians have started to use PDAs in an interactive way, that is, to write prescriptions, to keep a record of all daily clinical patient interactions, and for bedside charting. This trend was confirmed by another study by Martin (2003), who noted that more then 50% of physicians working in developed countries and under the age of 35 used a PDA in 2003.

Similar findings were also reported in other studies. In a related survey, it was found that between 40-50% of all U.S. physicians (including junior doctors, i.e., residents) were either using a PDA or had the ability and the knowledge to use a PDA in healthcare settings (Miller, Hillman, & Given, 2004).

The above findings have been supported by a detailed study by Carroll and Christakis (2004) who noted that from a random study of 2130 physicians, about 35-40% of physicians were using a PDA and that the most common use (80%) of PDAs was for drug referencing. The recent advances in RFID technologies (RFID Journal, 2005) are further promoting the adoption of mobile computing technologies (e.g., PDA) among healthcare stakeholders who have to deal with patient information (i.e., medication, allergies, etc.) on the location of an incident (Dwivedi, Bali et al., 2007; Meletis et al., 2006).

EMERGENCE OF RFID AS THE CRITICAL FACTOR FOR SUPERIOR HEALTHCARE DELIVERY

The origins of Radio Frequency Identification (RFID) systems can be traced to the emergence and consequent large scale adoption of automatic identification procedures (Auto-ID) in purchasing and distribution logistics, and manufacturing companies. The original objective of automatic identification procedures was to provide detailed information about people, goods, and products during transit, so as to enable better tracking. Barcode labels, which today are omnipresent, were the first automatic identification procedures that triggered a revolution in identification systems. Recently, due to their low storage capacity and the fact that they cannot be reprogrammed, they have become less popular, despite being very cheap (Finkenzeller, 2003).

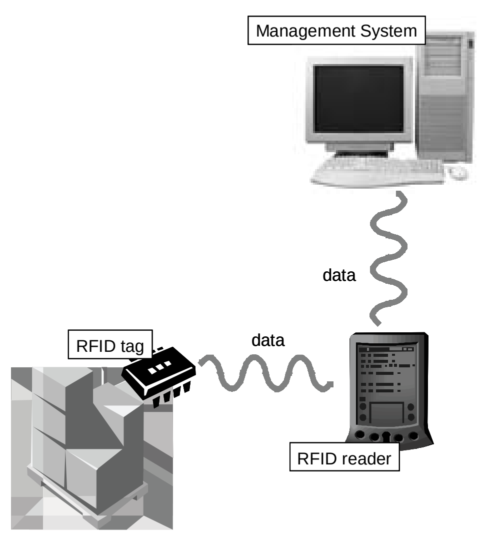
The rapid advances in computer hardware technology enabled the storage of data in a silicon chip. A disadvantage of this system was that smart cards, in order to be read, had to be placed in a smart card reader, which would in turn supply the smart card with power and a clock pulse, which was necessary for the smart card to function. However, the key drawbacks of smart cards as an automatic identification procedure are (a) smart card readers are expensive to maintain, (b) they tend malfunction, and (c) smart card readers kept in public places are prone to acts to vandalism (Finkenzeller, 2003).

These drawbacks in turn led to the evolution of automatic identification procedures wherein ideally the power required to operate the electronic data-carrying device was to be transferred from the reader using contact-less technology, and due to this connotation contact-less automatic identification procedures are called RFID systems (Finkenzeller, 2003).

An RFID as demonstrated in Figure 1 consists of three components: (1) a tag, (2) a reader, and (3) a computer network. The tag consists of a microchip that has some specific data that enables easy identification and an antenna, which is used to enable transmission of data. The second component of a RFID system is a reader and the reader uses radio waves to read the tag, and transmit data a computer system (i.e., the third component of a RFID system). The computer system enables the processing of information (Davis, 2004; Fanberg, 2004; Wilding & Delgado, 2004). This information system is illustrated in Figure 1.

RFID systems enable automatic identification and detail location of physical goods. Individual items or batches of goods carry an RFID transponder or "tag" that transmits a radio frequency signal. This signal can be remotely detected by an RFID "reader." When connected to a materials management system, the data downloaded from the reader is used to monitor and control the movement of those goods (Davis, 2004; Fanberg, 2004; Wilding & Delgado, 2004). In fact, it is the remote communication capability of RFID which differentiates it from existing traceability technologies. Previous automatic identification technologies such as printed batch cards and bar coding required individuals to read or scan the item/batch specific data at the location of the goods. Previous automatic identification

Figure 1. Radio Frequency Identification System (Davis, 2004; Fanberg, 2004; Wilding & Delgado, 2004)



technologies are also time consuming, laborious, and prone to inaccuracies, due to the scale and complexity of typical warehousing and distribution operations.

A number of studies (Anonymous, 2006a, 2006b, 2006c; Barlow, 2006; Carroll & Christakis, 2004; Davis, 2004; Fanberg, 2004; Hoppszallern, 2006; Pitts, 2005; Schuerenberg, 2005; Scott, 2006; Wicks, Visich, & Li, 2007; Williamson, 2006) have recommended the use of RFID in a healthcare context. In October 2005, the need to incorporate RFID in a healthcare context gained a big push when the Food and Drug Administration (FDA)—the U.S. based regulatory body—approved the use of RFID devices for medical purposes in humans, and listed RFID devices as a Class II medical device with special controls (Schuerenberg, 2005).

Some of the key advantages to be gained by the adoption of RFID in a healthcare context include: (1) tracking and combating counterfeit drugs (Fanberg, 2004; Pitts, 2005), (2) obtaining critical patient medical information, when the patient is clinically unable to provide the same (Schuerenberg, 2005), (3) preventing medication errors, (Anonymous, 2006b; Schuerenberg, 2005; Wicks et al., 2007), (4) tracking medical equipment (Anonymous, 2006c; Briggs & Martin, 2006; Scott, 2006), and (5) reducing surgical errors (Williamson, 2006).

RFID HEALTHCARE SUCCESS STORIES

St. Clair Hospital, a 331-bed hospital based in Pittsburgh, Pennsylvania, was keen to improve patient safety by employing IT to prevent errors in medication administration. In 2004, the hospital introduced a medication verification system that combined PDAs with scanning devices. The scanning devices would fit in the PC card slot on the PDAs. The nurses would use the PDAs to scan bar codes located on: (a) patient wristbands, (b) medications, and (c) their name/identification cards. When the system was introduced, nurses discovered that the entire process of scanning increased the time they had to spend on the process of medication administration (Schuerenberg, 2005).

Subsequent analyses revealed that a large part of the increase could be traced to the fact that the hospital's software required nurses to log in to the hospital system either by scanning the bar code on their ID badge or typing in a password on the PDA via a virtual keyboard. Furthermore, the use of bar codes meant that the nurses had to position the PDA near a patient's bar coded wristband to scan it and as their and/or patients wristbands became worn out, it would take the nurses several attempts before they could get a good scan on older bar coded wristbands (Schuerenberg, 2005).

To resolve this problem, the hospital administrators replaced the bar coded wristbands, patient wristbands, and clinical identification badges with RFID tags and the scanning devices that would fit into the PC card slot on the PDAs with a RFID tag. The introduction of RFID tags significantly reduced the time spent by the nurses on medication administration whilst also ensuring that medication errors were reduced. In addition, as RFID devices can exchange information without the need for direct device to device contact, as is the case for bar coded wristbands, the patient would not have to disturbed during the process, which was a regular occurrence when the bar coded wrist band was attached to the patient's arm (Schuerenberg, 2005).

Another use of RFID devices in healthcare has been to track clinical devices. In the U.S., as per Joint Commission on Accreditation of Healthcare Organization requirements, hospitals are responsible for ensuring that all their devices are regularly maintained or upgraded (as the case may be). Hospital staff have often complained that they have trouble tracking down medical devices (e.g., ventilators, intravenous pumps, etc.), some of which are regarded as life-saving-medical devices. The problem also has an administrative perspective as they have to tracked in-time for maintenance purposes as per the Joint Commission on Accreditation of Healthcare Organization requirements (Scott, 2006).

The paediatric critical care unit at the Vanderbilt Children's Hospital in Nashville, Tennessee, was concerned about the fact that, when required, a large part of their medical equipment could not be located, and often a search for medical equipment would result in location of their equipment in other parts of the hospital facility. When medical equipment could not be found, rental expenses for replacement equipment would cost the between $3,000 and $6,000 a month (Davis, 2004).

Vanderbilt Children's Hospital commissioned American Biomedical Group Inc. (ABGI), an Oklahoma City-based technology management company to implement a RFID solution with a deadline of 6 months for implementation to tackle their problem at a cost of $500,000. During the project, ABGI discovered that less than 50% of the paediatric critical care unit medical equipment were available at any given time, and adoption of RFID managed to alleviate their problems of equipment tracking (Davis, 2004).

Christiana Hospital, based in Newark, New Jersey, was concerned about the staff time lost and the distress caused in the time taken to locate patients who were based in the hospitals emergency rooms. A study in November 2004 found that "multiple phone calls and walking tours had to be carried out" in the hospitals emergency department to locate about 20% of their admitted patients. The hospital was able to resolve this problem by adopting a RFID system and within a 12 month period the hospital noticed dramatic improvements along a number of measurable targets (Briggs & Martin, 2006).

Mercy and Unity Hospital, a U.S.-based hospital was concerned about the effectiveness of their security mechanisms for tracking patient suffering from Alzheimer's Disease. In a pilot scheme, they used RFID tags which were embedded on patient wristbands to track patients suffering from Alzheimer's Disease, so as to prevent them from leaving the hospital unit. During the 4-month test period, adoption of RFID equipped wristbands helped reduce the number of patient watches by about 60% and saved about $30, 000 in costs (Davis, 2004).

The U.S.-based Joint Commission on Accreditation of Healthcare Organizations (JCAHO) has noted that it receives about five to eight new voluntary reports of surgical errors each month and further suspects that more cases of surgical errors are not being reported. The JCAHO further revealed that 76% of voluntary reports of surgical errors involved surgery on the wrong body part or site (Williamson, 2006).

In September 2005, Richard Choie, a physician at Washington School of Medicine, St. Louis, Missouri, adapted RFID technology to prevent surgical errors, and has termed the system as CheckSite. The objective of this system is to enforce surgical site marking that in turn will prevent surgical errors. The system has been implemented at Barnes Jewish Hospital in St. Louis in the U.S. and consequent to its adoption, "there have been no wrong site surgeries and no near misses" (Williamson, 2006).

The CheckSite system requires the physician to mark the surgical site on the patients body—this is done in consultation with the patient or the patient's representative. After the site is marked, the physician will place a special sticker on the patient's wristband, which has a RFID chip which will be deactivated by the sticker. If this process is not done before the surgery, the wristband will emit visual or auditory signals, and will page the concerned hospital personnel, thereby preventing surgical errors. This system is also financially quite cheap, as it costs only $2.50 per patient, apart from some installation expenditure (Williamson, 2006).

A more expensive and robust application of RFID technology to prevent surgical errors is the SurgiChip. It is the world's first RFID product approved for marking an anatomical surgical location. The SurgiChip system initially embeds surgical related patient information on an RFID smart label and on a chip that travels with the patient into surgery to help prevent errors. The information is also placed in the patient's file. Before the operation, the information on the chip is confirmed with the patient's chart and ID wristband (Williamson, 2006)

In February of 2004, the FDA, a U.S. based regulatory body had issued an guideline to pharmacy manufacturers that required them to provide a "pedigree" for each of their products. The origins of the term pedigree came from the art world wherein pedigrees are accepted as proof of authentication (Fanberg, 2004).

GlaxoSmithKline (GSK), in 2006, commenced an innovative RFID project to provide a pedigree for Trizivir (an HIV medicine) so as to enhance patient safety. GSK have started to put RFID tags on all bottles of Trizivir distributed in the U.S. These tags when scanned at a close range verify that the medication contained in the bottle is Trizivir. This mechanism was considered necessary as the U.S.-based National Association of Boards of Pharmacy had listed Trizivir as one of 32 drugs most susceptible to counterfeiting and diversion medication. Currently this project has cost GSK several million dollars and it has allowed GSK to track the genuineness (i.e., pedigree) of the medication as it moves through the distribution chain unto the at the point of dispensation to the consumer (Anonymous, 2006a)

The cost of implementing an RFID system in healthcare ranges from about $20,000 to $1 million—the variance dependent upon the size of the area where RFID technology is to be deployed and the application (e.g., patient tracking, equipment tracking etc) for which it is required. A typical return on investment on a RFID in healthcare system typically is less than a year, and in exceptional cases, the annual ROI can be up to 450% a year (Davis, 2004).

FUTURE OF RFID

Recent estimates on adoption of RFID systems in healthcare organizations indicate that RFID adoption in healthcare is on the verge of a dramatic rise beginning 2007. This estimate is partly based on a survey of more than 300 respondents. The results from the survey indicated that about 75% of respondents noted that adoption of RFID would lead to significant improvement in to patient safety and this was regarded to be the prime mover behind the projected rise of RFID adoption in healthcare (Anonymous, 2006b). The contention that RFID adoption in healthcare is on the verge of a dramatic rise is also supported by another study which has noted that, despite the fact that currently only 4% of U.S. hospitals use RFID to track moveable equipment, a recent study in 2005 by Hospitals & Health Networks Most Wired Survey indicated that over the next 2 to 3 years almost all hospitals will be adopting RFID based solutions to track medical equipment (Scott, 2006).

In 2006, the FDA commented that it is disappointed with the slow progress shown by drug manufacturers in adopting RFID solutions to enable better tracking of drugs so as to prevent counterfeiting. Since then, the FDA has released a draft compliance policy guide for public comment which makes a statement of origin of prescription drugs, that is, their "pedigree," compulsory (Anonymous, 2006b).

Another key obstacle in the uptake of RFID adoption in healthcare is lack of clear industry or government guidance on standards for adoption of RFID technology in Healthcare and about 60% of respondents indicated that they will delay their decisions on RFID adoption until clear guidelines or industry practices emerged. (Anonymous, 2006b).

CONCLUSION

On December 22, 2005, John Halamka, M.D., and CIO at Boston's CareGroup Healthcare System was the first person in the world to have an RFID chip embedded on his shoulder for medical use. The chip, whose size is about the size of a grain, was implanted in a painless 15 minute medical procedure under local anaesthesia, and it was inserted in a two-inch area on the arm between the elbow and shoulder. The chip inserted was manufactured by Applied Digital and is known as VeriChip. It contained a 16-digit identification number. The RFID chip can be read via a Pocket Reader (one has to wave the RFID reader over the patient to have access to all of the patient's medical information available on Boston's CareGroup Healthcare System (Schuerenberg, 2005). The drawback is that if John Halamka is being treated in any other then Boston's CareGroup Healthcare System, the physicians would not have access to John Halamka's medical data.

We believe that in the near future, the John Halamka experience will become the norm. This is based on the success stories of RFID in healthcare and of RFID in other sectors, which clearly demonstrate that there is a clear advantage of using RFID, in both clinical and nonclinical settings. Whilst the use of RFID is not compulsory, unlike legislation for bar coding, RFID has clear advantages (e.g., scan multiple items at the same time, remote scanning, scanned items do not have to be in sight at the time of scanning, etc.) over the traditional bar code as an automatic ID procedure. However, more work is required: (a) to improve the interface and interconnection abilities of RFID with mobile computing devices like handheld PDAs or tablet computers, and (b) to enhance the ability of RFID devices to work seamlessly with mobile computing devices in accessing EPRs.

REFERENCES

Anonymous. (2006a). GlaxoSmithKline starts RFID anti-counterfeit tests. *Healthcare Purchasing News, 30*(5), 8.

Anonymous. (2006b). Healthcare sector survey shows RFID's popularity. *Healthcare Purchasing News, 30*(1), 8.

Anonymous. (2006c). RFID tracking app helps reduce stays. *Health Data Management, 14*(1), 35.

Barlow, R. D. (2006). Bar coding has earned its stripes. *Healthcare Purchasing News, 30*(6), 22-23.

Briggs, B., & Martin, Z. (2006). Christiana Hospital honored for groundbreaking project. *Health Data Management, 14*(3), 24-26, 28, 30, 32, 34, 36.

Carroll, A. E., & Christakis, D. A. (2004). Pediatricians' use of and attitudes about personal digital assistants. *Pediatrics, 113*(2), 238-242.

Davis, M., & Klein, J. (2000, February). Net holds breakthrough solutions. *Modern Healthcare,* 14.

Davis, S. (2004). Tagging along. *Health Facilities Management, 17*(12), 20-24.

Dwivedi, A., Bali, R. K., James, A. E., & Naguib, R. N. G. (2001). *Telehealth systems: Considering knowledge management and ICT issues.* Paper presented at the 23rd Annual International Conference of the IEEE - Engineering in Medicine and Biology Society (EMBS), Istanbul, Turkey.

Dwivedi, A. N., Bali, R. K., & Naguib, R. N. G. (2005). Knowledge management for healthcare: Using information and communication technologies for decision making. In M. Jennex (Ed.), *Case studies in knowledge management* (pp. 328-343). Hershey, PA: Idea Group.

Dwivedi, A. N., Bali, R. K., & Naguib, R. N. G. (2007). Telemedicine - the next healthcare delivery medium: Fad or Future? [Special issue on the future of technology in healthcare]. *International Journal of Healthcare Technology and Management, 8*(3-4), 226-249.

Dwivedi, A. N., Wickramasinghe, N., Bali, R. K., Naguib, R. N. G., & Goldberg, S. (2007). Critical success factors for achieving superior m-health success. *International Journal of Electronic Healthcare, 3*(2), 261-278.

Fanberg, H. (2004). The RFID revolution. *Marketing Health Services, 24*(3), 43-44.

Finkenzeller, K. (2003). *RFID handbook: Fundamentals and applications in contactless smart cards and identification* (2nd ed.). Chichester, England; Hoboken, NJ: Wiley.

Hoppszallern, S. (2006). Tracking for time savings. *Hospitals & Health Networks, 5*(4), 44.

Kennedy, M. (1995). Integration fever. *Computerworld, 29*(14), 81-83.

Latamore, G. B. (1999). Workflow tools cut costs for high quality care. *Health Management Technology, 20*(4), 32-33.

Martin, S. (2003). More than half of MDs under age 35 now using PDAs. *Canadian Medical Association Journal, 169*(9), 952.

McCormick, J. (1999). Wireless hospitals: New wave in healthcare technology. *Health Management Technology, 20*(6), 12-13.

Meletis, B. A., Dwivedi, A. N., Gritzalis, S., Bali, R. K., & Naguib, R. N. G. (2006). Providing secure m-access to medical information [Special issue on integrating mobility into the healthcare sector: The next generation of mobile health applications. *International Journal of Electronic Healthcare, 3*(1), 51-71.

Miller, R. H., Hillman, J. M., & Given, R. S. (2004). Physician use of IT: Results from the Deloitte research survey. *Journal of Healthcare Information Management, 18*(1), 72-80.

Morrissey, J. (2000). Internet dominates providers' line of sight. *Modern Healthcare, 30*(15), 72-92.

Morrissey, J. (2001). Wanting more from information technology. *Modern Healthcare, 31*(6), 66-84.

Morrissey, J. (2002). High on tech, low on budget. *Modern Healthcare, 32*(4), 57-72.

Pitts, P. (2005). Identifying counterfeit drugs. *Canadian Healthcare Manager, 12*(8), 33.

RFID Journal. (2005). *RFID and emerging technologies guide to healthcare report*. Retrieved February 14, 2008, from http://www.rfidjournal.com

Schuerenberg, B. K. (2005). Implantable RFID chip takes root in CIO...literally. *Health Data Management, 13*(3), 14, 20, 22.

Scott, M. (2006). So that's where it is. *Hospitals & Health Networks, 80*(4), 28, 30.

Serb, C. (2002). Healthcare at your fingertips? *Hospitals & Health Networks, 76*(1), 44-46.

Wicks, A. M., Visich, J. K., & Li, S. (2007). Radio frequency identification applications in healthcare. *International Journal of Healthcare Technology and Management, 7*(6), 522-540.

Wilding, R., & Delgado, T. (2004). The story so far: RFID demystified. *Logistics & Transport Focus, 6*(3), 26-31.

Williamson, J. E. (2006). Surgical errors: New products, protocols help slash the risks. *Healthcare Purchasing News, 3*0(1), 22-24.

KEY TERMS

Bar Codes: A barcode can be defined as data that is recorded in a form that is machine-readable, and can be used by machine barcode readers, and are typically used to implement Auto ID Data Capture systems.

FDA: Food and Drug Administration is a governmental agency within the United States Department of Health and Human Services. It consists of eight offices: (1) Center for Biologics Evaluation and Research (CBER), (2) Center for Devices and Radiological Health (CDRH), (3) Center for Drug Evaluation and Research (CDER), (4) Center for Food Safety and Applied Nutrition (CFSAN), (5) Center for Veterinary Medicine (CVM), (6) National Center for Toxicological Research (NCTR), (7) Office of the Commissioner (OC), and (8) Office of Regulatory Affairs (ORA).

Radio Frequency Identification: Radio-frequency identification (RFID) is an automatic identification method which uses devices called RFID tags. An RFID system consists of three components: (1) a tag, (2) a reader, and (3) a computer network. The RFID tag consists of a microchip that has some specific data that enables easy identification and an antenna, which is used to enable transmission of data. The second component of a RFID system is a reader and the reader uses radio waves to read the tag, and transmit data a computer system (i.e., the third component of a RFID system).

RFID Tagging of Pharmaceuticals

David C. Wyld
Southeastern Louisiana University, USA

INTRODUCTION

Pharmaceuticals play a more prominent role in American health care than in any other nation. The North American market today comprises 47% of the global prescription drug market, which now exceeds half a trillion dollars, with Americans spending approximately $251.8 billion annually on pharmaceuticals. This is up significantly from a decade earlier, when American consumption represented approximately one-third of the world market (IMS Health, 2006a). America's insatiable demand for prescription drugs has led to serious cracks in the drug supply chain of the world's leading pharmaceutical market.

This chapter examines the size and scope of the problem of counterfeit pharmaceuticals, both globally and in the United States. It looks at the impact this crisis is having both on public health and the pharmaceutical industry. Today, leaders in both the pharmaceutical industry and government are looking to more stringent regulations and radio frequency identification (RFID) tagging of pharmaceuticals as a way to head off the problems associated with an increasingly leaky drug supply chain. This chapter outlines the steps being taken in the United States to help secure this vital supply chain. Finally, an analysis is given of the impact the shift to electronic pedigrees of pharmaceuticals will have both on the pharmaceutical and RFID market spaces.

BACKGROUND

The World Health Organization (WHO) estimates that as much as 10% of the global pharmaceutical market—a half-trillion-dollar marketplace—is counterfeit. In some countries, the WHO estimates that 25% or more of the entire drug supply is counterfeit. The New York City-based Center for Medicines in the Public Interest recently predicted that by 2010, counterfeit drug sales will reach $75 billion worldwide, almost doubling from the estimated counterfeit sales in 2005. The Federal Bureau of Investigation (FBI) estimates that the financial impact of counterfeit drugs on U.S. companies is $30 billion a year (Brooks, 2006; Eban, 2006).

Today, the toll of counterfeit drugs is mounting worldwide. Consider these recent news headlines from around the globe:

- In Hamilton, Ontario, Canada, Abadir Nasr, a registered pharmacist, was arrested by the Royal Canadian Mounted Police. Mr. Nasr, who was working at a retail drug store, dispensed counterfeit doses of Norvasc® to heart patients—pills filled with only talc. The local coroner investigated five patient deaths—all caused by a heart attack or stroke—that may have been brought about by the substitution of the counterfeit drug (Pitts, 2005).
- Within the last year, counterfeit versions of three popular drugs—Lipitor® for cholesterol, Cialis® for erectile dysfunction, and Reductil® for obesity—have surfaced in England. One British expert, Graham Satchwell, has estimated that 100,000 counterfeit drug imports are dispensed by the U.K.'s National Health Service annually (Eban, 2006).
- In China, a counterfeit drug smuggling ring was recently broken up that involved almost a half-million fake pills, including Lipitor® and all three major erectile dysfunction drugs (Cialis®, Levitra®, and Viagra®). Eleven Chinese nationals were arrested in the scheme, along with one U.S. citizen. The American had only recently been released from a New York State penitentiary for his involvement with counterfeit drugs in his home country, and he was drawn to the lucrative China market upon his release from jail due to the considerable money involved.
- Lipitor® is a cholesterol-reducing medication taken by more than 600,000 Americans, making it the most widely prescribed drug in the country. The U.S. Food and Drug Administration (FDA) announced on August 31, 2005, that it had busted a Lipitor® counterfeiting and smuggling ring that

was trafficking almost $50 million worth of the drug (Gottlieb, 2005).

There is already evidence that counterfeit drugs are worsening public health in general around the globe. For instance, in many Southeast Asian countries, more than half of the drugs sold do not have the correct formulation or levels of the active ingredient (National Drug Intelligence Center, 2005). This is especially telling in the case of malaria. Deaths from the dreaded disease are on the rise, despite the introduction of the latest and most effective antimalarial drug, Artemisinin®. Because of the poor quality of Artemisinin—both legitimate and illegitimate outlets—drug resistance to antimalarials is rapidly increasing. In fact, it has been estimated that more than 100,000 people die annually directly as a result of counterfeit antimalarial medications in the region (Stevens, 2006). Likewise, new drug-resistant forms of the HIV virus that causes AIDS have been largely attributed to the use of counterfeit medications for HIV-positive individuals (Gaul & Flaherty, 2003). In the United States, customs agents recently seized a large quantity of Tamiflu®, which is highly sought after today due to fears of avian flu. Epidemiologists fear that fake Tamiflu® could actually exacerbate any HN51 outbreak, as the counterfeit treatment could actually help the virus mutate into new drug-resistant strains (Stevens, 2006).

Scott Gottlieb, M.D. (2005), the FDA's Deputy Commissioner for Medical and Scientific Affairs, recently outlined the negative health outcomes that may occur directly from the use of counterfeit pharmaceuticals. These include actual cases where the fake drugs have contained ingredients that were:

- Inactive
- Incorrect
- Improperly dosed
- Subpotent
- Superpotent
- Expired
- Contaminated

Gottlieb (2005) summarized, "The result is risks to patients' health, either risk to their safety directly if the products are dangerous or risks from people suffering from complications from the many diseases that prescription drugs can treat today—but that the counterfeit versions cannot."

Why is there such unfortunate growth in counterfeit pharmaceuticals? The answer is multifaceted and complex, but the causal forces can be captured in the following:

- Profitability of the activity
- Relative ease of the activity
- Demand for drug products
- Cost of prescription drugs
- Web of country-specific regulations
- Vast cost disparities among countries on products
- Ease of transporting pharmaceuticals (which are generally shipped in cases, not pallets)
- Practice of relabeling, repackaging, and reimporting controlled substances
- Low prospect of being caught once the counterfeit pharmaceuticals are integrated into the drug supply (Pitts, 2005).

Indeed, according to the U.S. Department of Justice, the lure of counterfeit pharmaceuticals is so enticing that both organized crime and rogue entrepreneurs around the world are increasingly turning to the production and trade of fake, legal pharmaceuticals over narcotics and other illegal drugs (Gaul & Flaherty, 2003). The global pharmaceutical industry's counterfeiting problem is only exacerbated by the nature of its supply chain. In fact, the industry has been characterized in a recent *CIO Magazine* article as having "one of the world's most complex and opaque supply chains," producing "a web of legitimate, quasi-legitimate and illegitimate trade" (Patton, 2006, n.p.). With that, the sourcing of counterfeit drugs is now global, as fake pharmaceuticals have come from not only China and India but also from Central and South American as well as African and European countries (Stevens, 2006).

The United States has long thought of itself as being immune from the type of counterfeit drug problems found "over there" in areas of the world such as Southeast Asia and Africa (Pitts, 2005). This is because in the United States, 90% of all prescription drugs pass through the systems of just three drug wholesalers on their way to retail and hospital pharmacies. These are:

- AmerisourceBergen
- Cardinal Health
- McKesson (Navas, 2005)

As Gaul and Flaherty (2003) described the situation, pharmaceutical executives, retail pharmacies, and government regulators have relied upon these consolidated intermediaries as providing a "straightforward chain" that is the "gold standard" for control over the drug supply.

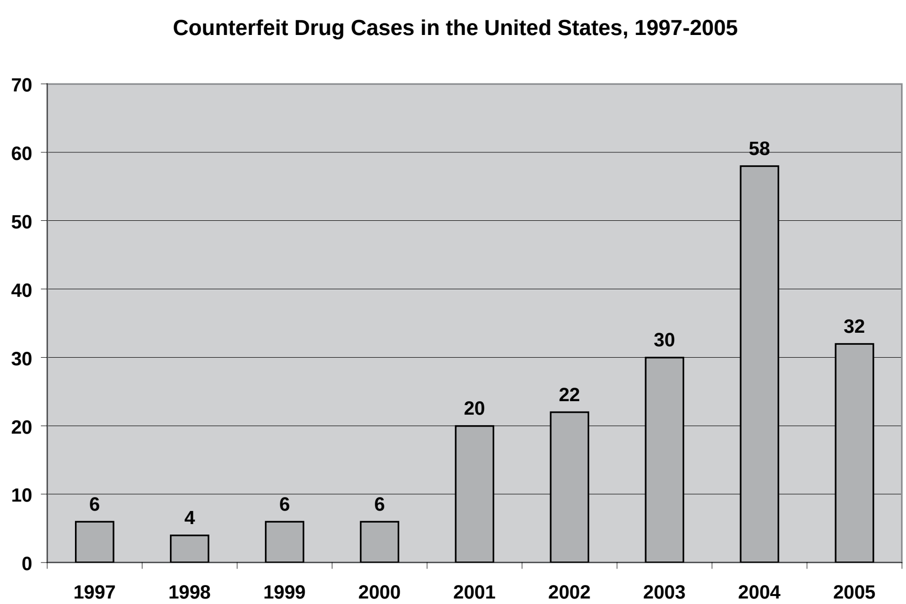
According to the FDA, the number of actual prosecuted cases of counterfeit pharmaceuticals has risen dramatically in the past decade. As can be seen in Figure 1, while there was a slight decline last year, which has been attributed to the increased focus of law enforcement authorities on the problem of counterfeit drugs, the number of cases this decade is far above historical levels (Brooks, 2006). In fact, the FDA's investigations of counterfeit medicines have increased by *almost tenfold* since 2000. Even the FDA's most conservative estimate—that approximately 1% of our nation's drug supply is counterfeit—amounts to some 35 million prescriptions a year. While this is far below the global rate of 10% estimated by the WHO, it is still an astonishing figure (Eban, 2006).

What drugs are being counterfeited? In the United States, the National Association of Boards of Pharmacy (NABP) coordinates the National Specified List of Susceptible Products for use by those in the pharmaceutical industry and law enforcement. The NABP specifies that for a drug to be included on its counterfeit watch list, it must meet one or more of the following criteria:

- Drug is prescribed for serious or life-threatening conditions
- Injectable drug or immune globulin has a single source
- Product has already been adulterated, counterfeited, diverted, or stolen
- Price is high
- Pedigree has been faked or not provided (Top 32 drugs most susceptible to counterfeiting, 2006).

With the latest addition of Viagra®, the NABP's (2004) list (see Table 1) has now reached 32 medications. As can be seen, these susceptible prescription medications include everything from cancer and hypertension drugs, drugs for those fighting HIV, and the "usual suspects" of erectile dysfunction medications. According to the most recent market statistics, Americans take more than 100 million prescriptions of these 32 drugs combined, which represent nearly $30 billion in annual sales, or more than 10% of pharmacy revenues. In fact, four of the top 10 prescribed medications (see Table 2) are among the medications listed on the NABP's counterfeit watch list. Sales of

Figure 1. Counterfeit drug cases in the United States, 1997-2005



Source: U.S. Food and Drug Administration (FDA)

Table 1. The 32 prescription drugs most susceptible to counterfeiting

1.	Combivir® (lamivudine/zidovudine)
2.	Crixivan® (indinavir)
3.	Diflucan® (fluconazole)
4.	Epivir® (lamivudine)
5.	Epogen® (epoetin alfa)
6.	Gamimune® (globulin, immune)
7.	Gammagard® (globulin, immune)
8.	Immune globulin
9.	Lamisil® (terbinafine)
10.	Lipitor® (atorvastatin)
11.	Lupron® (leuprolide)
12.	Neupogen® (filgrastim)
13.	Nutropin AQ® (somatropin, E-coli derived)
14.	Panglobulin® (globulin, immune)
15.	Procrit® (epoetin alfa)
16.	Retrovir® (zidovudine)
17.	Risperdal® (risperidone)
18.	Rocephin® (ceftriaxone)
19.	Serostim® (somatropin, mannalian derived)
20.	Sustiva® (efavirenz)
21.	Trizivir® (abacavir/lamivudine/zidovudine)
22.	Venoglobulin® (globulin, immune)
23.	Viagra® (sildenafil)
24.	Videx® (didanosine)
25.	Viracept® (nelfinavir)
26.	Viramune® (nevirapine)
27.	Zerit® (stavudine)
28.	Ziagen® (abacavir)
29.	Zocor® (simvastatin)
30.	Zofran® (ondansetron)
31.	Zoladex® (goserelin)
32.	Zyprexa® (olanzapine)

Source: The National Association of Boards of Pharmacy (NABP)

Table 2. Top-selling U.S. pharmaceuticals, 2005

Rank	Product	U.S. Sales (U.S. $Billions)*
1	**Lipitor®**	**$8.4**
2	**Zocor®**	**4.4**
3	Nexium®	4.4
4	Prevacid®	3.8
5	Advair® Diskus	3.6
6	Plavix®	3.5
7	Zoloft®	3.1
8	**Epogen®**	**3.0**
9	**Procrit®**	**3.0**
10	Aranesp®	2.8

Source: IMS Health, January 2006

these four prescription drugs alone (Lipitor®, Zocor®, Epogen®, and Procrit®) reached $19 billion in 2005 (IMS Health, 2006b).

Yet, with the growing demand for prescription drugs, there is increasing concern among pharmaceutical and retail executives, along with government regulators and law enforcement authorities, over the rapid increase in the illicit prescription drug trade, with a significant rise in the theft and diversion of pharmaceuticals, particularly in the realm of narcotics. According to the most recent statistics available from the U.S. Department of Justice's National Drug Intelligence Center (NDIC) in its *National Drug Threat Assessment 2005*, there has been a 16% increase in such activity involving narcotic drugs in the 2000-2003 time period. In the United States, the rates of abuse for prescription drugs, particularly in the areas of painkillers and amphetamines, have increased sharply since the early to mid-1990s (NDIC, 2005).

Most of the theft of controlled substances is internal, as employees of the pharmaceutical drug manufacturers and commercial distributors account for most of the diverted pharmaceuticals (NDIC, 2005). Likewise, according to the most recent National Retail Security Survey report, retailers attribute 48% of their pharmaceutical inventory loss to employee theft. The remainder of pharmaceutical theft comes from shoplifting, break-ins and armed robberies of pharmacies and clinics, and theft from individuals, most commonly from friends or relatives who possess legitimate prescriptions for the desired medications (Hollinger & Dabney, 2002).

All told, according to a confidential report from the Pharmaceutical Security Institute (PSI), a research organization funded by the pharmaceutical industry, the United States experienced 76 incidents of counterfeiting, diversion, and theft of pharmaceuticals in 2004. To put this number in perspective, according to PSI, this is the highest reported number in the world, ahead of even China and Colombia (Eban, 2005).

Counterfeit drugs cost pharmaceutical companies an estimated $46 billion annually (Patton, 2006). Considering the fact that pharmaceutical companies expend sometimes hundreds of millions of dollars to develop new drugs, this is a hit directly on their profits. According to the analysis of Professor David Taylor (2005) of the University of London School of Pharmacy, "The average selling price of a patented medicine is many times the cost of its basic ingredients. This is because "sunk" costs associated with research, manufacturing and marketing must be recouped within a narrow window of time" (n.p.). In fact, the marginal production cost of a new medicine will by definition be much closer to its ingredient cost and its eventual generic price. Just making one additional batch of a product will typically require very little extra spending. Thus, counterfeit drugs take away incremental revenue from the pharmaceutical companies, the vast majority of which would have gone straight to their bottom lines.

In actuality however, counterfeit drugs can have an even more dramatic impact. Bret Kinsella of ODIN Technologies stated, "In their most benign manifestation, counterfeit drugs are akin to placebos" (quoted in Wyld, 2006b, p. 44). Yet their very presence in the market weakens legitimate brands; when patients and doctors fail to see results from the use of what are unknowingly counterfeit products, they will be less likely to use or prescribe that brand of drug in the future. Thus, counterfeit drugs do not simply take away from a fixed pie of sales for a particular brand or even an entire product line, as over the long term, counterfeits can weaken or even kill a brand's and even a company's long-term image and prospects (Harrop, 2006).

RFID TAGGING OF PHARMACEUTICALS

There have been significant developments in regards to RFID tagging of prescription drugs. Leading players in the pharmaceutical marketspace have recently announced significant RFID labeling programs for three of the most sought after (both legally and illegally) drug products:

- Pfizer announced that by the end of the first quarter of 2006, all shipments of Viagra®, the lifestyle drug for erectile dysfunction, would carry RFID tags (Gohring, 2006). Likewise, the major drug wholesaler McKesson will use Viagra® as the subject of its own RFID trial (Patton, 2006).
- The privately held Purdue Pharma has greenlighted a significant pilot program to apply smart labels to all shipments of OxyContin®, a narcotic painkiller (Wyld, 2006a).
- GlaxoSmithKline (GSK) has begun tagging its all bottles of its Trizivir®, a drug for HIV infection (Trizivir bottles get ID tags, 2006).

For Pfizer, Purdue Pharma, and GSK, the decisions to move to RFID on these specific items were based on the popularity of these products, both in the mainstream and grey markets. All are among the leading counterfeited and diverted prescription drugs today. For GSK, the choice of Trizivir was due to the fact that it was found to be one of the 32 most counterfeited drugs. According to Peggy Staver, Pfizer's Director of Trade Product Integrity, the decision to implement RFID was an easy one for her firm, commenting, "It was an easy decision for us, as it's safe to say that Viagra has been our most counterfeited item" (Smith, 2006, n.p.). Likewise, Purdue Pharma is concentrating its RFID-tagging efforts on shipments of OxyContin®. According to Aaron Graham, Purdue Pharma's Vice President of Corporate Security, RFID gives his firm new visibility and control over the sensitive supply chain for OxyContin®, which he describes as a "a high-flyer outside of legitimate commerce" (Wyld, 2006a, n.p.)

There have been significant developments in the area of RFID labeling of pharmaceuticals, as Pfizer, Purdue Pharma, and GlaxoSmithKline have all announced major RFID tagging programs of their most counterfeited and diverted products in the early part of 2006 (Wyld, 2006b). Despite these recent announcements from industry leading firms, both the federal and state governments are initiating new regulatory efforts to tighten the pharmaceutical supply chain.

At the state level, attention has focused on assuring the pedigree of the drugs being given to patients. Several states, including Florida and California, are beginning to enact their own regulations over the pharmaceutical supply chain (Skrinar, 2005). Such laws call for an "e-pedigree," providing the ability to track all controlled substances from the manufacturer to the wholesaler's distribution center and ultimately to the pharmacy (Wyld, 2005). A number of states, including Arizona, Indiana, Oklahoma, and Texas, are instead focusing on requiring at least paper pedigree information, not on all prescription drugs but rather those medications that are acquired from outside of what is being called the "normal distribution channel" (Wasserman, 2005). As can be seen in Figure 2, this would generally focus on the third model of pharmaceutical distribution, where drugs come from outside the country or from atypical sources. Still, such pedigree requirements will not necessarily necessitate RFID, as most states' requirements can be met through the use of paper records and bar codes (ABI Research, 2006).

At the federal level, the FDA has been consistent in its position that better control over prescription drugs is one of today's most pressing public health and safety concerns. In a February 2004 policy statement, the FDA (2004a) advised that "use of mass serialization to uniquely identify all drug products intended for use in the United States is the single most powerful tool available to secure the U.S. drug supply" (n.p.). The FDA recommended that the pharmaceutical industry move to implement RFID tagging throughout the supply chain for controlled substances by 2007. While at present the FDA has not moved to mandate the usage of RFID at the individual item level, it has strongly encouraged all participants in the drug supply chain to cooperatively work to make this possible. To that end, in November 2004, the FDA (2004b), wanting to keep the agency's oversight function to a minimum, issued a directive that allowed pharmaceutical companies to implement RFID labeling and product tracking without having to first gain approval from the agency.

In the end, could RFID be made mandatory in the labeling of prescription drugs in the world's largest pharmaceutical consuming country—the United States? If some members of the U.S. Congress have their way, what has been advice and wishes from the federal agency overseeing the American pharmaceutical sector could become federal law. There are two bills before Congress that would lead to such an RFID mandate (Roberti, 2006).

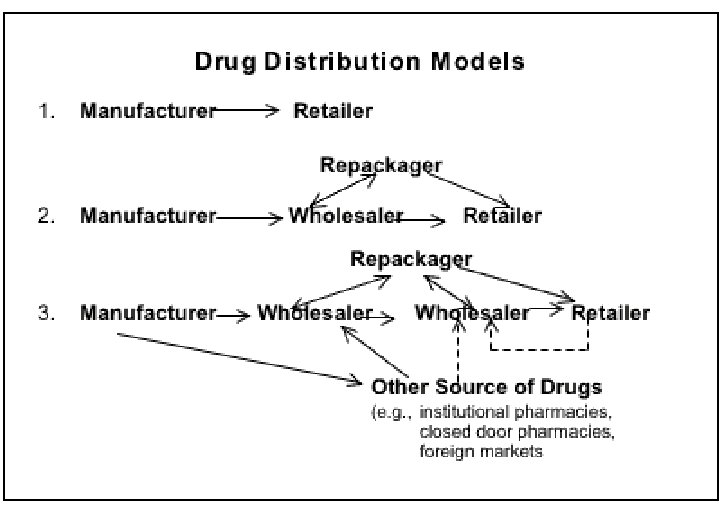
FUTURE TRENDS

There are exciting opportunities for research in the use of RFID in the pharmaceutical market. Looking ahead, researchers in the areas of management, marketing, supply chain management, logistics and distribution, and operations management will have chances to conduct a variety of firm-specific investigations, including:

- Case studies on RFID implementations
- Analyses of best practices
- Operations/economic impact

There will also be interest in longitudinal studies, looking at the impact of the introduction of RFID technology on the protection of the drug supply from counterfeiting. Similar studies could compare the RFID penetration in the pharmaceutical market to other retail sectors.

Figure 2. Drug distribution models



Source: U.S. Food and Drug Administration, FDA's Counterfeit Drug Task Force Interim Report, October 2003

CONCLUSION

Pharmaceuticals are indeed likely to be one of the first consumer-level applications of RFID tagging of products. With all the concern over counterfeiting drugs among both regulators and the pharmaceutical companies, one would expect to see only rosy forecasts for RFID in the pharmaceutical sector. However, two recent reports have taken starkly contrary positions on just how fast RFID-based labeling and control will spread in the pharmaceutical marketspace.

IDTechEx issued a report—RFID Forecasts, Players & Opportunities: 2006-2016—which was extremely bullish on the pharmaceutical RFID market. As shown in Table 3, IDTechEx predicts that the tag market alone will grow by more than 700% in the next year alone and grow almost exponentially over the next five years as more companies choose to—or are mandated to—tag pharmaceuticals at the item level in more areas of the global market. From the perspective of the report's author and IDTechEx's Chairman, Dr. Peter Harrop (2006), the rapid growth of RFID in the pharmaceutical sector can be explained by the "unusually broad range of benefits" item level tagging offers in this sector. He commented, "Frankly, no other form of RFID can claim such a full range of benefits, including saving lives, preventing sickness, reducing theft, fraud, counterfeiting and costs, providing more responsive customer service and recalls of higher integrity" (n.p.). Harrop (2006) predicts that "drug tagging has so many compelling drivers for all in the value chain as well as regulatory authorities, it may rise to be around 60% of all item level tagging in 2010" (n.p.).

On the bearish side, Oyster Bay, New York-based ABI Research warns of a slowdown in the move to RFID in this critical sector. According to ABI's 2006 report, titled "The RFID Healthcare and Pharmaceutical Markets," ABI Research (2006) warns that despite the success of Pfizer's and Purdue's highly publicized pilots, the pharmaceutical industry will only start tagging a total of about 10 pharmaceutical products by the end of this year. Sara Shah, ABI Research's industry analyst for RFID, goes so far as to observe that the earlier forecasts for the speedy implementation of RFID in the pharmaceutical sector may have been a case of "irrational exuberance" (ABI Research, 2006, n.p.).

Whichever trend line proves to be correct, the important point is that RFID does provide compelling benefits for the pharmaceutical sector in its ability to head off the explosion of counterfeiting going on today. By 2010, with or without government mandates, the pharmaceutical sector will, of necessity, move to RFID labeling of its products. If not, multibillion dollar investments in life-saving medicines could be evaporated overnight, and worse, the lives of billions of people may lie in the balance as our aging and growing populations

Table 3. Predictions for item-level tagging of pharmaceuticals over the next decade

Year	Regions Affected	Activity Level	Approximate Number of Tags (in Millions)
2006	• USA	• Some prescription drugs in the USA	35
2007	• USA • Europe	• Some prescription drugs in the USA	270
2008	• USA • Europe • East Asia	• Many prescription drugs in the USA	1083
2009	• USA • Europe • East Asia	• All prescription drugs in the USA • Many prescription drugs abroad	5250
2010	• USA • Europe • East Asia	• All prescription drugs in the U.S. • Many prescription drugs abroad • Possibly some OTC (over-the-counter) nonprescription medications	1600

Source: IDTechEx, RFID Forecasts, Players & Opportunities, 2006-2016 depend on the ingenuity—and genuineness—of the pharmaceuticals that keep us alive.

REFERENCES

ABI Research. (2006). *Legislation uncertainty slows adoption of pharmaceutical RFID tracking: The RFID healthcare and pharmaceutical markets*. ABI Research, February 6, 2006. Retrieved April 1, 2006, from http://www.abiresearch.com/abiprdisplay.jsp?pressid=594

Brooks, K. (2006). *Anti-counterfeiting initiatives and RFID practices*. Contract Pharma, January 2006. Retrieved April 2, 2006, from http://www.contractpharma.com/articles/2006/01/anticounterfeiting-initiatives-and-rfid-practices.php

CVS pulls out of secondary drug market. (2005). *Pharmacy Times*, September 2005. Retrieved February 20, 2006, from http://www.pharmacytimes.com/Article.cfm?Menu=1&ID=2562

Eban, K. (2005). How counterfeiters are contaminating America's drug supply: The problem of tainted medicine in America. *Dangerous Doses*, May 2005. Retrieved January 15, 2006, from http://www.dangerousdoses.com/pdf/scope.pdf

Eban, K. (2006). Where good and bad drugs mix: Why drug importation poses real risks for American consumers. *Medical Progress Today*, March 16, 2006. Retrieved April 2, 2006, from http://www.medicalprogresstoday.com/spotlight/spotlight_indarchive.php?id=1176

Gaul, G.M., & Flaherty, M.P. (2003). U.S. prescription drug system under attack: Multibillion-dollar shadow market is growing stronger. *The Washington Post*, October 19, 2003. Retrieved February 6, 2006, from http://www.washingtonpost.com/ac2/wp-dyn/A44908-2003Oct18?language=printer

Gohring, N. (2006). Pfizer to tag Viagra with RFID to control fakes. *BIO-IT World*, January 10, 2006. Retrieved February 25, 2006, from http://www.bio-itworld.com/newsitems/2006/january/01-10-06-viagra/view

Gottlieb, S. (2005). *Ensuring the safety of America's drug supply*. Proceedings of the Drugs Counterfeiting Conference of the Center for Medicines in the Public Interest. Retrieved March 25, 2006, from http://www.cmpi.org/media/Drugs_Counterfeiting_papers.pdf

Harrop, P. (2006). Pharmaceutical RFID—Fast forward: The requirements of the pharmaceutical industry are very different from those of any other industry and this is reflected in its rapid adoption of RFID. *IDTechEx*, January 26, 2006. Retrieved February 2, 2006, from http://www.idtechex.com/products/en/articles/00000415.asp

Hollinger, R.C., & Dabney, D.A. (2002). Social factors associated with pharmacists' unauthorized use of mind-altering prescription medications. *Journal of Drug Issues, 31*(1), 231–264.

IMS Health. (2006a). *IMS Health reports 5.4 percent dollar growth in 2005 U.S. prescription sales: New product launches, Medicare Part D help fuel projected U.S. pharma market growth of 5-to-8 percent over next five years*. Retrieved May 1, 2006, from http://www.imshealth.com/ims/portal/front/articleC/0,2777,6599_3665_77180090,00.html

IMS Health (2006b). *U.S. top-line industry data, 2005*. Retrieved May 1, 2006, from http://www.imshealth.com/ims/portal/front/indexC/0,2773,6599_73914113_0,00.html

National Association of Boards of Pharmacy (NABP). (2004). *National specified list of susceptible products*. Retrieved April 20, 2006, from http://www.nabp.net/ftpfiles/NABP01/List.pdf

National Drug Intelligence Center (NDIC). (2005). *National drug threat assessment 2005*. Retrieved March 14, 2006, from http://www.usdoj.gov/ndic/pubs11/12620/pharma.htm

Navas, D. (2005). Visibility along the pharma supply chain. *Supply Chain Systems Magazine*. Retrieved July 7, 2006, from http://www.scs-mag.com/index.php?option=com_content&task=view&id=358&Itemid=90

O'Connor, M.C. (2006). FDA to update its RFID vision. *RFID Journal*. Retrieved February 20, 2006, from http://www.rfidjournal.com/article/articleprint/2148/-1/1/

Patton, S. (2006). Cracks in the pharmaceutical supply chain: Faced with a rising tide of counterfeit and mispriced drugs, pharmaceutical companies are turning to technologies such as RFID to better track medications through a convoluted supply chain. *CIO Magazine*. Retrieved May 14, 2006, from http://www.cio.com/archive/011506/pharma.html

Pitts, P. (2005). *21st century terrorism*. Proceedings of the Drugs Counterfeiting Conference of the Center for Medicines in the Public Interest. Retrieved March 25, 2006, from http://www.cmpi.org/media/Drugs_Counterfeiting_papers.pdf

Roberti, M. (2006). Congress weighs drug anticounterfeiting bill: Recently introduced legislation would compel the FDA to require drugmakers to incorporate counterfeit-resistant technologies, such as RFID, into the packaging of prescription drugs. *RFID Journal*. Retrieved April 10, 2006, from http://www.rfidjournal.com/article/articleprint/2180/-1/1/

Skrinar, T. (2005). Securing the extended supply chain: Looking for clarity in the RFID rush. *Pharmaceutical Business Strategies, 1*, 10.

Smith, A. (2006). *Keeping tabs on Viagra et al: Looking to stop theft, drug industry could switch to RFID within five years*. CNNMoney.com, January 26, 2006. Retrieved March 20, 2006, from http://money.cnn.com/2006/01/26/news/companies/RFID/

Stevens, P. (2006). Counterfeit drugs: Getting to the root of the problem. *Medical Progress Today*. Retrieved April 20, 2006, from http://www.medicalprogresstoday.com/spotlight/spotlight_indarchive.php?id=1182

Taylor, D. (2005). *Private concerns and public interests*. Proceedings of to the Drugs Counterfeiting Conference of the Center for Medicines in the Public Interest. Retrieved March 25, 2006, from http://www.cmpi.org/media/Drugs_Counterfeiting_papers.pdf

Top 32 drugs most susceptible to counterfeiting. (2006b). *Supply Scape*, January 2006. Retrieved May 15, 2006, from http://www.supplyscape.com/product/mostlikely.php

Trizivir bottles get ID tags. (2006a). *Drug Topics*, April 10, 2006. Retrieved April 30, 2006, from http://www.drugtopics.com/drugtopics/article/articleDetail.jsp?id=318509

U.S. Department of Health and Human Services, Food and Drug Administration (FDA) (2003). *FDA's Counterfeit Drug Task Force—Interim report*. Retrieved November 13, 2004, from http://www.fda.gov/oc/initiatives/counterfeit/report/interim_report.html#IIA

U.S. Department of Health and Human Services, Food and Drug Administration (2004a). *Combating counterfeit drugs: A report of the Food and Drug Administration—February 2004*. Retrieved November 15, 2004, from http://www.fda.gov/oc/initiatives/counterfeit/report02_04.html

U.S. Department of Health and Human Services, Food and Drug Administration (2004b). *Radiofrequency identification feasibility studies and pilot programs for drugs—Guidance for FDA staff and industry: Compliance policy guides, November 2004*. Retrieved November 22, 2004, from http://www.fda.gov/oc/initiatives/counterfeit/rfid_cpg.html

Wasserman, E. (2005). Purdue Pharma to run pedigree pilot: The prescription drug manufacturer will use RFID to track and authenticate bottles of OxyContin as they move through the supply chain—From factory to pharmacy. *RFID Journal*. Retrieved July 20, 2005, from http://www.rfidjournal.com/article/articleview/1626/1/1

Wyld, D.C. (2005). *RFID: The right frequency for government—A research monograph published by the IBM Center for the Business of Government*. Retrieved October 15, 2005, from http://www.businessofgovernment.org/main/publications/grant_reports/details/index.asp?gid=232

Wyld, D.C. (2006a). Viagra and Oxycontin tagged, but future still uncertain for RFID in pharma. *ID Magazine, 2*(2), 56–58.

Wyld, D.C. (2006b). Fatal fakes: A look at how counterfeit pharmaceuticals are becoming a major global public health crisis and a significant financial drain on the pharmaceutical industry. *Global Identification*, May, 42-45.

KEY TERMS

Active Tag: A type of RFID tag that has its own power supply (battery or external power), and, when interrogated by a reader, the tag emits its own signal. Typically, active tags have far greater read distances than passive tags, and they can be combined with sensors to provide information on the environment and condition of the item. They are also more expensive than passive tags and – due to the battery – have a limited life span.

Automatic Identification (Auto-ID): This is a broad term, encompassing technologies used to help machines identify objects. A host of technologies fall under the Automatic Identification umbrella, including bar codes, biometrics, smart cards, voice recognition and RFID.

Electronic Article Surveillance (EAS): Acknowledged by many as the first widespread use of RFID technology widely used today in the retail environment and in libraries, these systems use microwave or inductive technology "readers" to detect the presence or absence of EAS tags as a means of detecting and deterring theft. When an item is purchased (or borrowed from a library), the tag is turned off. However, when someone passes a gate area holding an item with an EAS tag that hasn't been turned off, an alarm sounds. These tags are inexpensive and do not contain any data.

Electronic Product Code: This is a unique number - stored in the chip on an RFID tag -that identifies an item in the supply chain, allowing for tracking of that item. EPC is the generally accepted acronym for "Electronic Product Code."

Frequency: The number of repetitions of a complete wave within one second. 1Hz equals one complete waveform in one second. 1KHz equals 1,000 waves in a second. RFID tags use low, high, ultra-high and microwave frequencies. Each frequency has advantages and disadvantages that make them more suitable for some applications than for others.

Harvesting: The way passive tags gather energy from an RFID reader's antenna to be able to respond to the reader.

Passive Tag: A type of RFID tag that does not have its own power supply. Instead, the tag draws power from the reader, which sends out electromagnetic waves that induce a current in the tag's antenna. Without an onboard power source, passive tags have a lesser read range than active tags. However, they cost less than active tags and have an unlimited life span.

Power Level: The amount of radio frequency (RF) energy radiated from a reader or an active tag. Higher power outputs enable longer read ranges. However, most governments regulate the power levels at which RFID readers can operate in order to avoid interference with other RF devices.

Radio Frequency Identification: An automatic identification technology that uses radio waves to identify objects. RFID is the generally accepted acronym for Radio Frequency Identification.

Read: The process of retrieving data stored on an RFID tag by sending electromagnetic waves to the tag and converting the radio waves the tag sends back into data.

Read Rate: The number of tags that can be read by an RFID reader in a given time period.

Read/Write: The ability of an RFID system to change the data that are stored in a tag. For example, as a product moves from the final packaging area to the warehouse, a read/write tag can be modified to reflect the new location so that now, when interrogated, it passes the new location as part of its updated data stream.

Reader (also called an interrogator): A device that communicates with RFID tags. The reader has one or more antennas, which emit radio waves and receive signals back from the tag. Readers may have a digital display to relay information to the operator and may transmit data on to an organization's computer network infrastructure. Readers can be either fixed or portable, and today, they are beginning to be integrated into other electronic devices such as PDAs (personal digital assistants) and cell phones, and even into objects such as pens.

Semi-Passive Tag (also called battery-assisted tags): This type of tag is similar to an active tag in that there is an onboard battery. The battery is used to run the microchip's circuitry and to boost the effective read range of the tag. Some semi-passive tags sleep until they are woken up by a signal from the reader, which conserves battery life, while some are programmed to broadcast at set intervals of time.

Sensor: A device that responds to a physical stimulus and produces an electronic signal reporting on that stimulus. Sensors can be tailored to report on a variety of environmental conditions, including temperature, movement, vibration, and shock.

Slap and Ship: A manufacturer placing RFID tags on cases and/or pallets at the last possible point before shipping from a supplier to a mandating retailer or other organization. With a slap and ship strategy, a manufacturer or distributor is simply trying to meet the requirements of another firm's RFID mandate rather than attempting to capture any data—and value—from the RFID tagging in their own system.

Smart Label: A generic term that refers to a printed label that typically contains printed information, a bar code identifier, and an RFID tag. The label is considered to be "smart" because of its ability to communicate with an RFID reader.

Robust Blood-Glucose Control of Type I Diabetes Patients Under Intensive Care Using Mathematica

Béla Paláncz
Budapest University of Technology and Economics, Hungary

Balázs Benyó
Budapest University of Technology and Economics, Hungary

Levente Kovács
Budapest University of Technology and Economics, Hungary

Zoltán Benyó
Budapest University of Technology and Economics, Hungary

INTRODUCTION

This article presents a robust control design method on frequency domain using *Mathematica* for regularization of glucose level in Type I diabetes persons under intensive care. The method originally proposed under *Mathematica* by Helton and Merino (1998) is now improved with a disturbance rejection constraint inequality, and is tested on the three-state minimal model. Nonlinear closed loop simulation in state-space, in case of standard meal disturbances, demonstrates the robustness of the resulted high-order linear controller. The obtained results are compared with H_∞ design implemented with Matlab, proving that the controller (for the considered model parameters) can operate properly, even in case of parameter values of the worst-case scenario.

The blood-glucose control is one of the most difficult control problems to be solved in biomedical engineering. The main reason is that patients are extremely diverse in their dynamics and, in addition, their characteristics are time-varying. The investigations of Hernjak and Doyle (2005) discourage the use of a low complexity control such as PID, if high level of performance is desired. To design an optimal, high-quality control, one needs a relevant model of the process, as well as a proper control technique. There are several studies in both areas (Parker, Doyle, & Peppas, 2001).

The mostly used model, and also the simplest one, proved to be the minimal model of Bergman, Philips, and Cobelli (1981), but its shortcoming is its big sensitivity to variance in the parameters. Henceforward, the plasma insulin concentration must be known as a function of time. Therefore, extensions of this minimal model have been proposed (e.g., de Gaetano and Arino (2000) analyzed the different models of glucose-insulin interactions in human body). Also, more general models have been used in the literature, like the model developed by Hovorka, Shojaee-Moradie, Carroll, Chassin, Gowrie, Jackson, Tudor, Umpleby, and Jones (2002), or the 21st order nonlinear model of Sorensen (1985).

However, probably the best way to approach the problem of blood-glucose control is to consider the system model and the applied control technique together (Makroglou, Li, & Kuang, 2006; Parker et al., 2001). As models of diabetic systems are imprecise by nature, some research works are concentrated on adaptive control techniques using online parameter estimation (e.g., Lin, Chase, Shaw, Doran, Hann, Robertson, Browne, Lotz, Wake, & Broughton, 2004), others like Ruiz-Velazquez, Femat, and Campos-Delgado (2004) or Parker, Doyle, Ward, and Peppas (2000) suggest robust control design.

This study focuses on robust control design, for regulating glucose level in Type I diabetes patients under intensive care. The investigations were done using the minimal patient model of Bergman et al. (1981). The method applied using *Mathematica* lays on a different approach by that under Matlab. The corresponding *Mathematica* program package OPTDesign was originally developed by Helton and Merino (1998). The technique of the computations and closed loop simulations, employing the *Mathematica* Application, Control System Professional Suite (CSPS), together with OPTDesign, has been already demonstrated by Palancz (2006).

Mathematica is a mathematical program capable to work step by step, like a mathematical reasoning machine. In this way, it offers a more didactic approach as Matlab, which is focused on solving the problems in a more technical way (Palancz, Benyo, & Kovacs, 2005).

Here the authors have slightly improved this technique by suggesting an effective disturbance rejection constraint inequality for the disturbance transfer function. In order to check the quality, especially the robustness, the controller is designed with the most favorable model parameter values, but tested with the values representing the worst-case scenario in terms of difficulty of the system dynamics for control purposes. Results are compared with those obtained by Hernjak and Doyle (2005).

MODEL EQUATIONS

The three-state minimal patient model of Bergman et al. (1981) consists of the following equations:

$$\begin{aligned}\dot{G}(t) &= -p_1 G(t) - (G(t)+G_B)X(t) + h(t)\\ \dot{X}(t) &= -p_2 X(t) + p_3 Y(t)\\ \dot{Y}(t) &= -p_4(Y(t)+Y_B) + i(t)/V_L\end{aligned} \quad (1)$$

where the three state variables (as well as outputs) are the plasma glucose deviation G(t) (mg/dL), remote compartment insulin utilization X(t) (1/min), and plasma insulin deviation Y(t) (mU/dL). The control variable is the exogenous insulin infusion rate, i(t) (mU/min), whereas the exogenous glucose infusion rate h(t) (mg/(dL min)) represents the disturbance.

Other variables represent parameters of system given by Equation 1. The physiological parameters are G_B (mg/dL) the basal glucose level, Y_B (mU/dL) basal insulin level, V_L (dL) the insulin distribution volume and p_1, p_2, p_3, p_4 as the model parameters. As numerical values, the authors worked with the numerical values determined by Furler, Kraegen, Smallwood, and Chisolm (1985): p_1 = 0.028 (1/min), p_2 = 0.025 (1/min), p_3 = 0.00013 (dL/(min^2 mU)), p_4 = 5/54 (1/min), G_B = 110 (mg/dL), Y_B = 1.5 (mU/dL), V_L = 120 (dL).

The steady-state values used for linearizing the system are: G0 = X0 = Y0 = 0, h0 = 0, and for i0:

$$i_0 = p_4 Y_B V_L = 16.667 \text{ (mU/min)} \quad (2)$$

Loading CSPS of *Mathematica* the linearized system around the vicinity of the steady-state can be calculated. The system proved to be stable, controllable, and observable.

CONCEPT OF THE ROBUST CONTROL DESIGN

Performance Requirements

Considering the complementary sensitivity function of a general closed loop system (Zhou, 1996):

$$T(s) = \frac{P(s)C(s)}{1+P(s)C(s)} \quad (3)$$

where P(s) represents the transfer function of the considered plant, and C(s) the transfer function of the controller. As a result, the robust control method on frequency domain implemented by OPTDesign briefly can be summarized, satisfying the following conditions (Helton & Merino, 1998):

1. T must satisfy disk inequality:

   $$|K(i\omega) - T(i\omega)| \le R(i\omega), \text{ for } \omega_a \le \omega \le \omega_b \quad (4)$$

   where K and R are fixed functions that embody the desired specifications of the system. K is called the center of the disk, and R is called the radius.
2. Defining the gain-phase margin as $m = \inf|1+PC|$, the constraint should be:

   $$|T(i\omega)-1| \le \frac{1}{m}, \text{ for all } \omega \quad (5)$$

3. The bandwidth of the complementary sensitivity function (T(iω)) should be below than 1/√2 or in other words below -3 dB (Zhou, 1996).
4. For the closed-loop roll-off, specifying a given n and α_r, as well as the roll-off frequency ω_r for which the $C(i\omega) \le \frac{\alpha_r}{|\omega|^n}$ inequality is held, then for large ω frequencies, it is true that $T(i\omega) \le |P(i\omega)C(i\omega)|$, or by other words:

$$|T(i\omega)| \leq \alpha_r \frac{|P(i\omega)|}{|\omega|^n}, \text{ for } |\omega| > \omega_r \quad (6)$$

In addition, the authors introduced a new condition for disturbance rejection requirement. Considering $P_d(s)$ the transfer function of the disturbance and having the sensitivity transfer function (1-T(s)), the inequality should be:

$$|1 - T(i\omega)| \leq \frac{c}{|P_d(i\omega)|} \quad (7)$$

where c is a constant less than 1.

These requirements all together can be summarized in the requirement envelope presented in Figure 1.

In order to ensure the control purposes, as well as the proper performance of the optimized process, the following performance requirements were chosen: $c = 0.95$; $\alpha_r = 8.7641*10^{-7}$; $n = 2$; $\omega_d = 2.65$; $\omega_b = 4$; $\omega_r = 6.5$; $\alpha_g = 2.5$; $\alpha_b = 0.9$. Results are shown in Figure 2.

Optimizing Performance Function

Assuming that T(s) has no right-half plane (RHP) poles (is internally stable), the designing problem can be formulated as it follows:

- Given a plant P(s), a set of performance requirements $\overline{P}$ and a set $\overline{I} = \{ T(s) \in RH^\infty \}$ of rational internally stable transfer functions the task is to determine optimal $T \in \overline{I}$ which satisfies $\overline{P}$.

Starting from Equation 4, the performance requirements can be expressed in form of disk inequalities, which can be written in the following way:

$$\frac{1}{R(i\omega)^2}(|K(i\omega) - T(i\omega)|)^2 \leq 1, \text{ for } \forall\omega \quad (8)$$

Now, for a given T, one can calculate the largest value of the left-hand side:

$$\gamma(T) = \sup_\omega \left(\frac{1}{R(i\omega)^2}(|K(i\omega) - T(i\omega)|)^2 \right) \quad (9)$$

If T is optimal (T^*), it means that $\gamma(T^*) \leq \gamma(T)$ for all $T \in \overline{I}$ and satisfying $\overline{P}$. By other words, this means:

$$\gamma^* = \inf_T \sup_\omega \left(\frac{1}{R(i\omega)^2}(|K(i\omega) - T(i\omega)|)^2 \right) \quad (10)$$

Figure 1. Requirement envelope for robust control design in frequency domain. Data from Helton and Merino (1998).

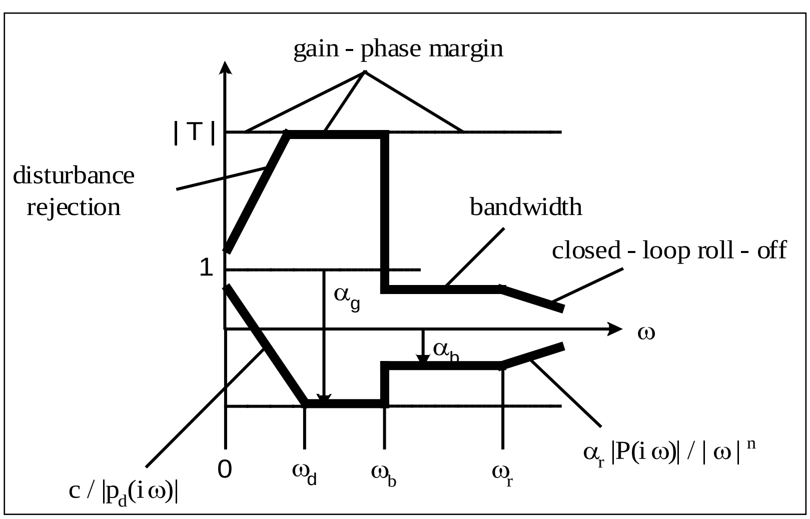
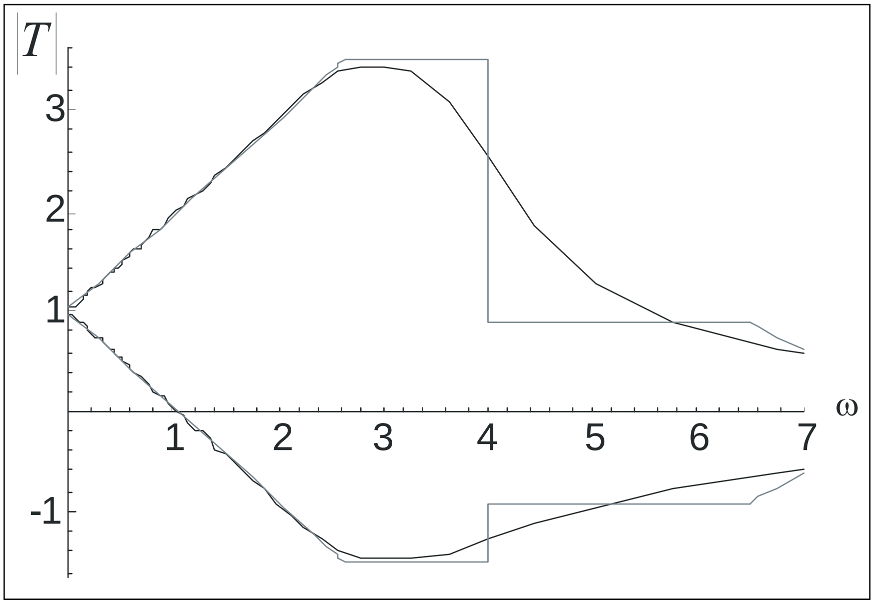


Figure 2. The actual original and smoothed requirements envelopes.



This γ^* value is basically the solution of the H_∞ suboptimal problem (Zhou, 1996) representing the worst-case performance of T^* (optimal worst-case performance of T). Its value should be in the [0,1] interval to guarantee robust performance (and also satisfying Equation 8). Furthermore, Helton and Merino (1998) demonstrated, there exist functions $A,\ B \in RH^\infty$ that:

$$T = A + BT^+, \text{ for some } T^+ \in RH^\infty \qquad (11)$$

or by other words, there exists a formula which gives all $T \in \bar{I}$ depending on RHP poles and zeros of the given plant P.

Running OPTDesign for the requirement envelope shown in Figure 2, the obtained result for γ^* was 0.3679. This value was also checked with the mu-toolbox of Matlab (working with a different approach of the H_∞ problem as *Mathematica* does), and a very similar result was obtained. Figure 3 presents the numerical values of the optimal T^* inside the considered constraining envelope. For internal stability parameterization the A and B functions were:

$$A(s) = \frac{1}{(s+2)^5},\quad B(s) = \frac{1}{(s+2)^5} \qquad (12)$$

With the calculated γ^*, the authors have checked the performance requirements similarly as in Palancz (2006), taking into account the specifications of Helton and Merino (1998). The resulted indicators diagnosing the optimality of the obtained solution are:

- Flat = 0.0032. It is a number between 0 and 1, measuring the "nonflatness" of the current guess. 0 represents the absolute optimal value.
- GrAlign = $1.9984 \cdot 10^{-15}$. It represents the gradient alignment, indicating how a certain winding number has the "appropriate value." Its true answer at optimality diagnostics is equal to 0.
- Ned = $4.35 \cdot 10^{-5}$. It is also a positive number, indicating when small that the numerical calculations contain little numerical error.

The numerical values of these optimal T^* will be approximated with a proper rational function. The minimal approximation error found a 20th order rational function. The local error of this approximation can be

Figure 3. The numerical values of the optimal T inside the considered constraining envelope.*

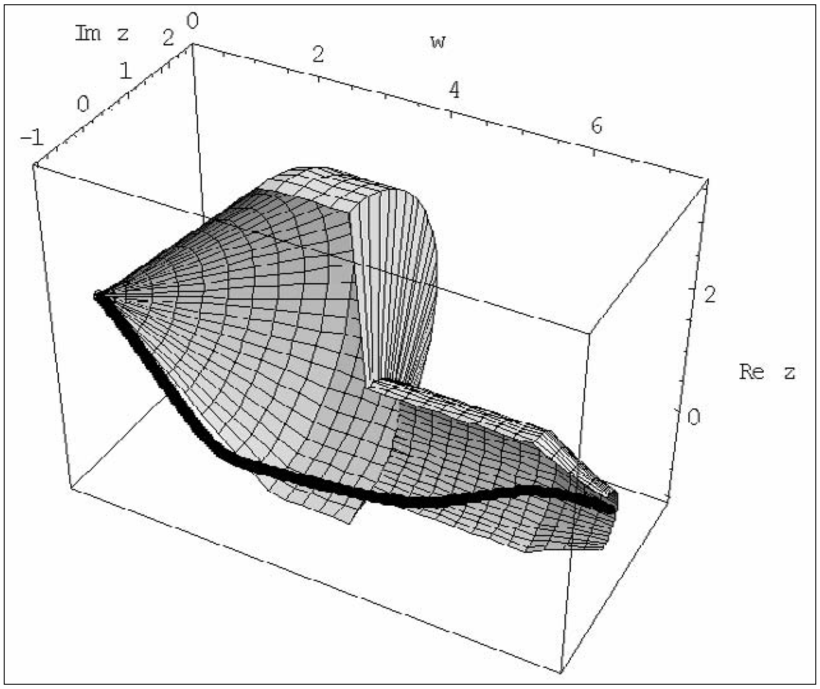
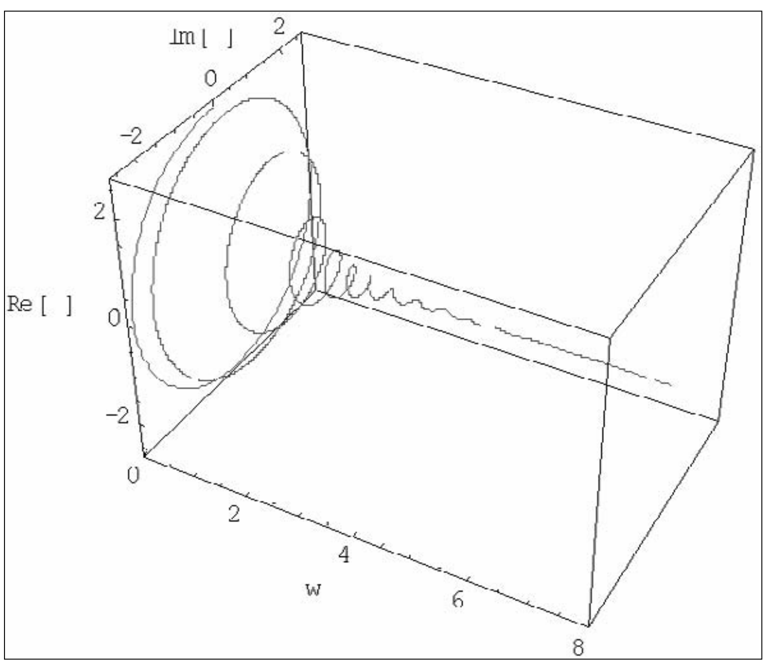


Figure 4. Local error of optimal rational T (in 10^{-4}).*



seen in Figure 4. Applying Equation 3, one can express the transfer function of the control part, C(s). In order to simulate the nonlinear closed-loop in time, domain C(s) should be converted in state-space form.

To reduce the size of the state-space form of the controller, the authors use the "MinimalRealization" technique built in CSPS. As a result, a sixth order model is obtained (see Box 1).

Nonlinear Closed Loop Model Simulation

Using the parameters given at the beginning of this article (see Equation 2), the performance of the control was tested by using a standard meal disturbance with about six hour duration. This was modeled by Lehman and Deutsch (1992) (see Figure 5).

Using the designed controller, Figure 6 and Figure 7 presents the controlled dynamics of the blood glucose and insulin infusion. These results are very similar to those published by Hernjak and Doyle (2005).

Testing Robustness of the Controller

The investigations of Hernjak and Doyle (2005) showed that considering the minimal model with $p_1 = 0$, the system is unable to regulate the glucose level on its own. Indeed, the eigenvalues of the linearized model in case p1 = 0 are {- 0.0925926, - 0.025, 0}. As a result, the open loop system is unstable, but controllable. The result of the simulation carried out with the same controller, but with nonlinear plant model having the

Box 1.

$$A=\begin{pmatrix} -5.488 & 0.189 & 0.0042 & -0.002 & -7.34\cdot10^{-6} & -1.27\cdot10^{-6} \\ -39.61 & -5.99 & -0.1335 & 0.066 & 2.29\cdot10^{-4} & 3.96\cdot10^{-5} \\ 81.03 & 12.05 & -1.754 & 0.876 & 0.003 & 5.23\cdot10^{-4} \\ 3823.16 & 568.03 & -83.024 & -5.837 & -0.02 & -0.003 \\ 50135.2 & 7448.91 & -1088.6 & -76.22 & -13.21 & -2.31 \\ -205374 & -30513.7 & 4459.35 & 312.25 & 54.54 & 5.26 \end{pmatrix}$$

$$B=\begin{pmatrix} -2.52\cdot10^{-5} & 1.87\cdot10^{-4} & 0.00038 & 0.018 & 0.237 & -0.97 \end{pmatrix}^T$$

$$C=\begin{pmatrix} -4.597\cdot10^{10} & 1.94\cdot10^{9} & 1.123\cdot10^{8} & -1.064\cdot10^{8} & 170668 & -33566 \end{pmatrix} \quad (13)$$

Figure 5. Considered exogenous glucose infusion (meal disturbance). Data taken from Lehman and Deutsch (1992).

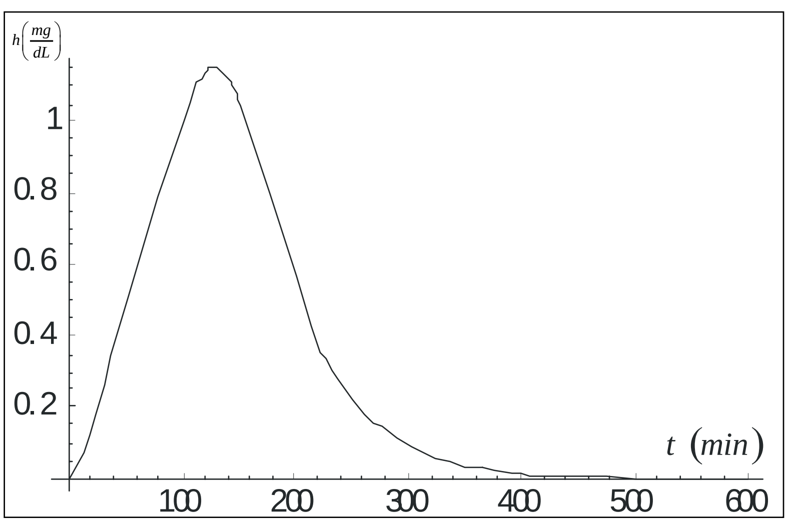


Figure 6. Controlled dynamics of blood glucose concentration, G(t)

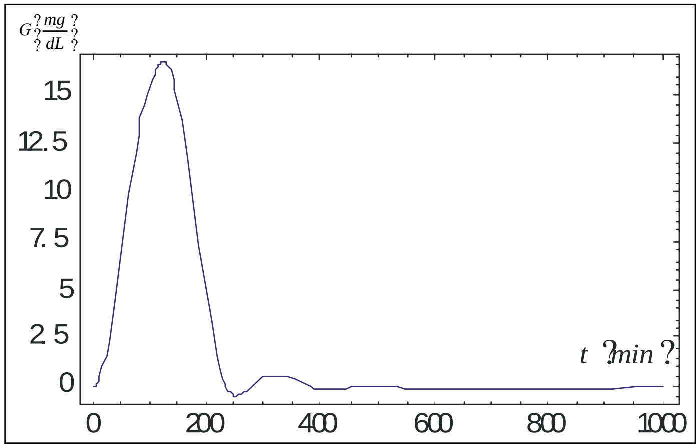
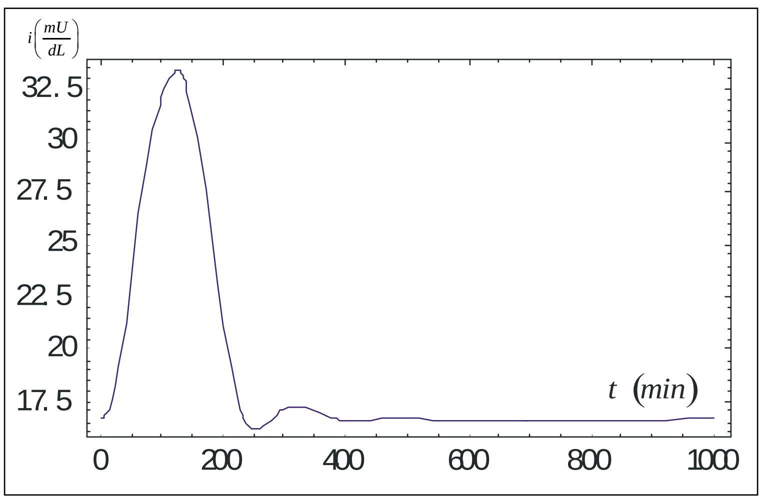


Figure 7. Controlled dynamics of insulin infusion rate, i(t).



value of p1 = 0, shows that the compensator is able to control the system even in this case. However, the quality of the control is not as good as it was in case of p1 = 0.028. Results are illustrated by Figure 8 and Figure 9, and these are very similar to those obtained by Hernjak and Doyle (2005) with applying NMPC (Nonlinear Model Predictive Control) control method (in case of NMPC the best result was obtained, and the controller was originally designed for instable situation, p1 = 0).

FUTURE TRENDS

In the future, investigations of presented robust control method using *Mathematica* will be tested for the minimal Bergman model extended with the mixed meal model (Roy & Parker, 2006). The results will be supported by clinical measurements.

From another point of view, the robust method will be tested on the Type I diabetes model of Sorensen (1985). In this case, the aim is also to reduce the model keeping its characteristics, but in a reduced way, one is able to investigate it much easier.

Figure 8. The controlled dynamics of blood glucose, G(t) in case of unstable plant, p_1=0.

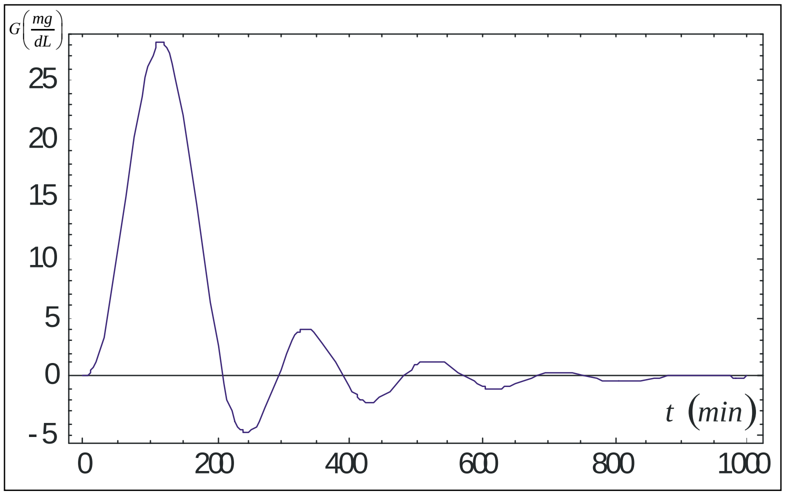
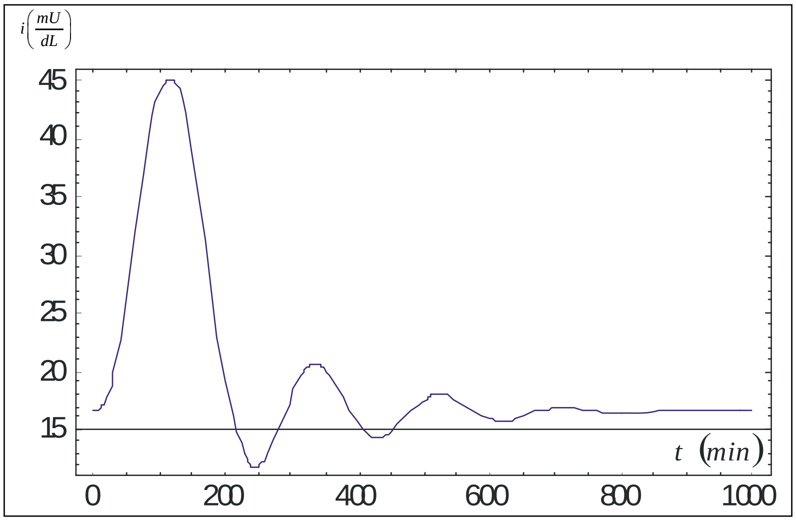


Figure 9. The controlled dynamics of insulin infusion rate, i(t) in case of p_1 = 0.



CONCLUSION

A practical method has been presented for designing a robust controller to regulate glucose-insulin system for Type I diabetic patients under intensive care. Employing *Mathematica* CSPS Application, together with the OPTDesign package developed by Helton and Merino (1998), one may define a linear, high-order compensator, in relatively easy way, which can be tested via nonlinear model simulation. Introducing a proper disk inequality constraint for disturbance rejection, this method is proved to be effective for providing acceptable control performance, even in a case when the model parameters have changed, and the system became unstable. After further verifications, this method could provide a useful help to control of blood glucose level, and in the optimization process of diabetic administration.

Furthermore, the proposed design procedure can be easily adapted to other problems too, where uncertainty plays an important role and robustness of the control is indispensable.

REFERENCES

Bergman, R. N., Philips, L. S., & Cobelli, C. (1981). Physiologic evaluation of factors controlling glucose tolerance in man. *Journal of Clinical Investigation*, *68*, 1456–1467.

de Gaetano, A., & Arino, O. (2000). Mathematical modeling of the intravenous glucose tolerance test. *Journal of Mathematical Biology*, *40*, 136–168.

Furler, S. M., Kraegen, E. W., Smallwood, R. H., & Chisolm, D. J. (1985). Blood glucose control by intermittent loop closure in the basal mode: Computer simulation studies with a diabetic model. *Diabetes Care*, *8*, 553–561.

Helton, J. W., & Merino, O. (1998). *Classical control using H_∞ methods, theory, optimization, and design*. Philadelphia: SIAM.

Hernjak, N., & Doyle, F. J. III. (2005). Glucose control design using nonlinearity assessment techniques. *AIChE Journal*, *51*(2), 544–554.

Hovorka, R., Shojaee-Moradie, F., Carroll, P. V., Chassin, L. J., Gowrie, I. J., Jackson, N. C., Tudor, R. S., Umpleby, A. M., & Jones, R. H. (2002). Partitioning glucose distribution/transport, disposal, and endogenous production during IVGTT. *American Journal of Physiology – Endocrinology and Metabolism*, *282*, 992–1007.

Lehman, E. D., & Deutsch, T. A. (1992). A physiological model of glucose - insulin interaction in type I diabetes mellitus. *Journal of Biomedical Engineering*, *14*, 235–242.

Lin, J., Chase, J. G., Shaw, G. M., Doran, C. V., Hann, C. E., Robertson, M. B., Browne, P. M., Lotz, T., Wake, G. C., & Broughton, B. (2004, September 1–5). Adaptive bolus-based set-point regulation of hyperglycemia in critical care. In *Proceedings of 26th Annual International Conference of the IEEE EMBS* (pp. 3463–3466). San Francisco, CA.

Makroglou, A., Li, J., & Kuang, Y. (2006). Mathematical models and software tools for the glucose - insulin regulatory system and diabetes: An overview. *Elsevier Applied Numerical Mathematics*, *56*(3–4), 559–573.

Palancz, B. (2006). H_∞ control design on frequency domain. *Wolfram Information Center*. Retrieved February 11, 2008, from http://library.wolfram.com/infocenter/MathSource/6013/

Palancz, B., Benyo, Z., & Kovacs, L. (2005). Control system professional suite. *IEEE Control System Magazine*, *25*(2), 67–75.

Parker, R. S., Doyle, F. J. III, & Peppas, N. A. (2001). The intravenous route to blood glucose control. *IEEE Engineering in Medicine and Biology*, *20*, 65–73.

Parker, R. S., Doyle, F. J. III, Ward, J. H., & Peppas, N. A. (2000). Robust H_∞ glucose control in diabetes using a physiological model. *AIChE Journal*, *46*(12), 2537–2549.

Roy, A., & Parker, R. S. (2006). Dynamic modeling of free fatty acid, glucose, and insulin: An extended "minimal model." *Diabetes Technology & Therapeutics*, *8*(6), 617–626.

Ruiz-Velazquez, E., Femat, R., & Campos-Delgado, D. U. (2004). Blood - glucose control for type I diabetes mellitus: A robust tracking H_∞ problem. *Elsevier Control Engineering Practice*, *12*, 1179–1195.

Sorensen, J. T. (1985). *A physiologic model of glucose metabolism in man and its use to design and assess improved insulin therapies for diabetes*. Unpublished doctoral thesis, Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge.

Zhou, K. (1996). *Robust and optimal control*. New Jersey: Prentice Hall.

KEY TERMS

Blood-Glucose Control: Control of blood sugar levels.

Control System Professional Suite (CSPS): Application in Mathematica.

Insulin: Hormone in the body.

Mathematica: A fully integrated software environment for technical and scientific computing.

Matlab: A tool for doing numerical computations with matrices and vectors.

The Role of Government and Its Influence on Nursing Systems by Means of the Definition of Nursing Minimum Data Sets (NMDS)

Bettina Staudinger
University for Health Sciences, Medical Informatics and Technology, Austria

Herwig Ostermann
University for Health Sciences, Medical Informatics and Technology, Austria

Magdalena Thöni
University for Health Sciences, Medical Informatics and Technology, Austria

Roland Staudinger
University for Health Sciences, Medical Informatics and Technology, Austria

INTRODUCTION

Since the 1980's, questions have been asked the world over about the efficiency and contribution of nursing in the hospital treatment of patients and nursing performance within the framework of nonhospital health care.

The cause for these tendencies has many roots. For one, we can determine a push in professionalism through the increasing importance of nursing sciences on whole. The basic focus is on comparability, standardization (Johnson et al., 2005), and securing quality (ICN, 2003). Also, a significant part of nursing systems internationally are publicly financed and legally determined. This has the consequence that the political decision-makers, particularly in context with the financing and planning of nursing structures, have more of an interest in controlling the nursing systems and disposing of useable nursing data.

The different health systems have reacted to this cost pressure, not least because of the various political conceptions according to national criteria, even though it has been established that the basic tendencies of individual steps to reform are comparable in all health systems. Because of this, a diversification of the nursing professions corresponding to national demands was noticed on an international scale and the attempt was made to assign the nursing performance to the level of training. At the same time, the structural, hierarchical cooperation of the different nursing professions was regulated.

A professionally and politically specific intention concerns the availability of valid nursing data which serve just as much as a political decision basis for developing the need for nursing performance in the individual professional levels as they serve as essential support within the framework of information of individual nursing processes. Initiatives for reform and the forming of data structures in nursing came from the nursing branch, or rather nursing sciences themselves, as well as from the government decision-makers.

Along with other reform objectives in the area of professional training and distinctions of competence, the focus is on the definition of a list of nursing performances (nursing interventions) in context with nursing data. These performances have to be clearly defined and serve as a basic fundament for nursing documentation as well as for care planning on a national scale.

Nursing Minimum Data Set (NMDS) has been developed as a scientific term for this process, although the term is understood in different ways in different countries.

BACKGROUND

Nursing as an independent profession has a more far-reaching tradition in America than it does in Europe. In Europe, nursing as an autonomous profession and also therefore the science of nursing, only established itself in the last 20 years of the 20th Century. It is therefore not surprising that many foundations and discussion anchors of nursing science and practical nursing have their roots in Anglo-American areas such as the USA or Australia. This also applies to a large degree to the discussion about minimum nursing data which were defined by American nursing scientists at the end of the 1980s (Nonn, Mayer, & Evers, 2002). While in

America the issue of recording systems has a prominent position, the discussion about minimum nursing data in Europe is accompanied by establishing a "professional identity," which allows it to be followed through the constitution of junior and superordinate nursing job descriptions with corresponding training levels.

Nursing sciences and political decision-makers function as communicating vessels during the development of national NMDS, even though the motives for this development are completely different.

Being responsible, politicians concern themselves with constructing and configuring the nursing the system itself; the nursing scientists are actually more occupied with the systematology and basic knowledge from the nursing data.

The intrinsic problem relating to the development of NMDS arises from the differing motives and information expectance. This is because the government as legislator and designer of the nursing system has to fall back on data structures which enable data consolidation on varying metalevels and therefore enforce and secure structural and financial of nursing data systems. In respect thereof, the structural elements are an important part of governmental considerations.

The interest of the nursing sciences is concentrated on the availability of data structures which allow outcome studies and other nursing studies on a scientific level being simultaneously suitable for the support of the individual nursing processes (Lamb, Mowinski Jennings, Mitchell, & Lang, 2004). During the development of national NMDS, this implicit discrepancy can be observed time and again by means of various process descriptions. For example, the description of the objectives of the Swiss NMDS project were at first fairly extensive and included not just the demands of nursing science but also showed clear interdisciplinary traits (NURSING Data Model, 2001). Whilst implementing the project itself, these demands had to be largely reduced and adapted (NURSING Data Final Report, 2004). However, the focus of governmental needs of data availability in order to target the nursing system was expanded upon quite considerably or remained unchanged in the core areas.

This conflict of interests during the development of the NMDS is on the whole distinctly recognizable in all national projects. There are, however, less changes to the initial objectives where the government took the initiative from the start to develop an NMDS and made its objectives an essential part of the NMDS project, a case in point being Australia. Here the National Minimum Data Sets are created, coordinated by the Australian Institute of Health and Welfare (AIHW) (2006), which is also responsible for a broad reporting system in health affairs.

Regarding this, it can be discovered that governmental motivation and position on developing a nursing system followed by the establishment of a NMDS is of high importance for the final concrete embodiment of an NMDS.

THE INFLUENCE OF THE GOVERNMENT ON THE NURSING SYSTEM AND THE DEVELOPMENT OF NMDS

Politics Influencing Nursing Systems

In current literature about nursing research, a lot of room is taken up by the question about how far political intervention and system parameters influence the nursing concept and therefore nursing outcome.

In contrast to English-speaking and Scandinavian countries, where the question about political influence and actions on the nursing system is well established in nursing research, this approach in the European-German speaking area has only just begun. Particularly in connection with the effect of changed quality criteria of a nursing unit, current studies based on the longitudinal analysis method, attempt to produce evidence—limited by political parameters and the structural changes in nursing associated with them—that a changing effect in quality of nursing was achieved or rather is verifiable (Blegen, Goode, & Reed, 1998).

The influence of politics on the nursing system and nursing outcome is shown as being particularly extensive in these studies. According to this trend, individual sections, for example, the question of mortality, are examined in connection with political system design. The results of American nursing research tend to show that a direct connection exists between nursing outcome—in this case mortality—and politically specified structures (Czaplinski & Diers, 1998).

Within nursing research, Oddi and Cassidy (1990) chose a systematic approach regarding the interaction between politics and nursing. Their approach assumes the protection of human rights by American law. Based on the "rights of human subjects" and the

"Patient's Bill of Rights," the authors categorize the rights of the patients on a "professional, public, and private" sphere.

Whilst the majority of studies have portrayed nursing as a monocausal passive system which is at the mercy of politics, and have simply referred to the effects of political intervention, the authors have rather assumed interaction. This is even more remarkable as the protection of the patient and his data from the researcher and misuse within the context of clinical research takes an important position. Here it is the task of the government to protect the patient through standards. The authors arrive at the conclusion: "As professionals, nurses must be actively involved in the formation of public policy…" and introduce a applicable model—over and above the nursing research—within the setting of American structures.

Gorenberg, Alderman, and Cruise (1991) also take the same line. They write: "As health care providers, nurses have a professional responsibility to act on behalf of the public they serve by influencing health policy."

In a more recent study, Maas and Delany take a completely different approach (Mass & Delany, 2004). The authors critically view the significance of nursing outcome studies and relate upon the quality of the studies, preliminary irrespective how far the political system intervention takes place or not. As a parameter, based on the approaches of Fawcett and Russell (2001) ("grounded in the social contract between nursing and society"), the authors discuss and examine which data the system currently supplies or is in the position to supply to nurses.

First of all, the quality of communication and documentation in American health care facilities is discussed and examined. The result being that a standardized "nursing language" associated with the nursing process is not evident in a quality that system related outcome studies or data based scientific evaluation requires. The authors refer to a series of current field studies and come to the conclusion that sufficient terminology was indeed developed for individual case-related nursing, but that this terminology is insufficient for analyses on a metalevel.

There actually exist various IT-supported documentation systems in American health care facilities, the focus of these systems, however, is geared and adapted to in-house processes; in context with the missing nursing terminology, a sufficient basis for data exploration can not be assumed. Altogether the authors assess that actually the theoretical and operative conditions are not eligible for outcome studies in the USA due to the single case orientation of the nursing system.

Hence for political intervention there is not sufficient basic data for political intervention concerning the nursing system and therefore—if nursing wants to achieve goal-orientated and system-supported system adjustment—these shortcomings have to be rectified. "National and international policy implications are also compelling. As noted, nursing practice data are not currently included in large national datasets […] are often used for quality analyses that influence health policies in the United States and other countries" (Marin, Rodrigues, Delaney et al., 2000).

Nursing research outside of the USA and GB is also concerned with the question of data quality in connection with political and systemic interventions. In particular, terminology questions in connection with the definition and documentation of the nursing process are the object of European and German research. Hinz and Dörre (2002) introduce one approach.

A central point here is also the development of comprehensive and exact nursing terminology based on the nursing diagnoses definitions and classification of the North American Nursing Diagnosis Association (NANDA). The further development of nursing terminology through the International Council of Nurses (ICN) represents, according to the authors' opinion, a useful instrument, even though it could be shown that these classification system and terminology are not sufficiently implemented into nursing practice in German-speaking areas (Schaubmayr, 2004).

The authors also assume IT support as the instrument to create the requirements for comparability, and consequently, systematic intervention which are goal-orientated and able to be evaluated.

Another important question in association with the effect of political intervention on the nursing system and consequently on nursing outcome, is the concept of politics in a narrower sense and the kind of political intervention that occurs.

Maas and Delaney (2004) use a very wide term of political structures, and consider particular institutions and legislators, superior control structures and management functions, as well as financiers of the health system in the wider sense as politically active institutions. Even if they do not define the interactions or the interdependences of the different control mechanisms for the nursing system in particular and even more do not go into the different appearance of control, this concept of policy and intervention corresponds to the described wide approach of system interaction.

National Projects for the Development of NMDS

The development of NMDS in the individual states exhibits not only different priorities, but also different stages of development. Whilst there is still no national NMDS project in Austria, Australia, for example, is at a very mature stage on a national level. Therefore, in this article chosen projects will be briefly described.

Since 1987, Belgium has made it obligatory in its general hospitals to collect nursing data according to the Belgian Nursing Minimum Data Set (BeNMDS). This NMDS was developed between 1983 and 1986 together with two university research centres and the Belgian nurse professionals association. The BeNMDS comprises of a list of twenty three nursing action related items (nursing interventions) which, together with encoded institution and patient data, is collected quarterly on 15 consecutive days. With this approach, Belgium dispensed with a part of the nursing process data, but the collection of nursing relevant data can serve as a basis for the calculation of nursing input and demand. Due to this data collection policy, Belgium is assumed to be the only country that has long term nursing data at its disposal for analysis and report.

Based on the construct of the Belgian Minimum Data Set, similar concepts were also developed in other European countries. The difference to Belgium is that not all countries implemented a simultaneous legal obligation for the collection of data, or rather nationwide collection of data. The Dutch Minimum Data Set (NeNMDS) which was developed in the 1990s comprises nursing interventions, nursing diagnoses, and nursing outcome categories (Goossen, 2001).

Along with proprietary development of NMDS, the BeNMDS was repeatedly tested and adapted for national requirements (Turtiainen, Kinnunen, Sermeus, & Nyberg, 2002), whereby the validity of the individual nursing interventions from the BeNMDS under the requirements of local expansion of the list of variables according to national requirements was asserted several times (Turtiainen et al., 2002; Volrathongchai, Delaney, & Phuphaibul, 2003).

An enormously extensive approach in nursing data collection was decided by the Australian government. One of the main differences lies in the fact that the data collection and definition of the Australian National Minimum Data Set (AuNaMDS) was not solely restricted to the hospital branch, but was in fact extended for all community care providers. For each of the different areas of medical care in the widest sense, minimum data sets were defined and from their basis, nationwide health statistics are derived. The development of these minimum data sets occurred in the 1990s and the objective was the collection of data for the planning of nationwide care. For this reason, no uniform Nursing Minimum Data Set exist, rather respective minimum data sets for different branches of care, for example, the minimum data set for Home and Community Care (HACC).

In Switzerland, the implemented project for the development of Swiss nursing data started in 1998 and is being carried out in several scheduled stages. The planning phase and the test phase have, according to reports from the specially founded organisation NURSING data, already been completed (Berthou, Junger, & Kossaibati, 2005). The acquired Swiss minimum data set (CH-NMDS), also contains data of care providing institutions and information about the careers (nurses), along with nursing phenomenon and nursing interventions, which are in their logical and terminological structure aligned with the WHO-ICF classification. What can also be seen as a characteristic is the range of the CH-NMDS, which is not just designed for the hospital sector, but should also be valid in Swiss home care and for nursing performance in nursing homes (Stark & Hölzer, 2005).

The Interaction of Government, Nursing Sciences, and Nursing Institutions

The demand for structuring and structured collection of nursing data is from the outset supported by the idea of data material for nursing research on the one hand and for the planning of required care of the population with nursing care requirements on the other. Furthermore, the nursing documentation should shape the provision of nursing care more transparently and efficiently, and through this have an advantageous effect on the care of the patient or client. For all fields of nursing care there is one common vital condition, namely the comparability of collected nursing data which in turn has the condition of the development of a common "generally applicable" nursing language.

This fundamental idea is also the starting point for consequently brisk development of nursing terminologies, or rather terminological systems which were developed for nursing and whose integration is still not completed. These systems have led to efforts for more extensive regional and content expansion.

What all terminology systems have in common is that they attempt to break down nursing actions into operating levels on the one hand and provide uniform practicable formulations for the elements of a nursing action on the other.

Running along side this, classification systems were created which process the individual steps of the nursing process in uniformly formulated catalogues. For the various steps of the nursing process, the respective classifications were compiled which have been coordinated by the individual research groups for several years in order to achieve complete terminological coverage of the nursing process.

NANDA-Nursing Diagnoses (NANDA International, 2005), the Nursing Interventions Classification (NIC, 1996), the Nursing Outcomes Classification NOC (Johnson et al., 2005), and the International Classification for Nursing Practice (ICNP®) all have their origins in the American nursing sciences and professional initiatives. They all follow different approaches for systemizing nursing language and the classification of nursing actions.

In Australia, terminological agreements had already been made through the above mentioned state initiatives and they were published in the "National Data Dictionaries" of the AIHW (Australian Institute for Health and Welfare).

Irrespective of origin, it can be assumed that electronic support of collection and documentation plays and intrinsic role in all initiatives for data collection in nursing and for extensive nursing documentation. This is ostensibly rooted in the content and procedural range of nursing. The scheduled evaluation of nursing data for different groups, for example, nurses themselves, nursing management, nursing scientists, nursing needs-planners and structural planners implicates an electronic availability of the data. This is confirmed by the fact that all initiatives for nursing data collection to date were and are accompanied by more or less extensive IT projects.

The interaction of government, nursing science, and nursing institutions in particular when defining a NMDS and irrespective of the given motivation is shown in Figure 1.

The common denominator preliminarily lies in the question of structuring and defining the systems and terminologies which is considered to be the basic requirement for the digitalization of the nursing process. Only from this digitalized nursing process data can be made both available and analyzable. From now on, the precise implementation of these data structures depicts the formed NMDS.

According to Werely, Devine, Zorn, Ryan, and Westra (1991), the basic definition of minimum nursing data is: "a minimum set of items [or elements] of information with uniform definitions and categories, concerning a specific aspect or dimension of the health care system, which meets the essential needs of multiple data users." Multiple data users implicates different action partners who have an interest in the implementation of nursing documentation but whose objectives are different.

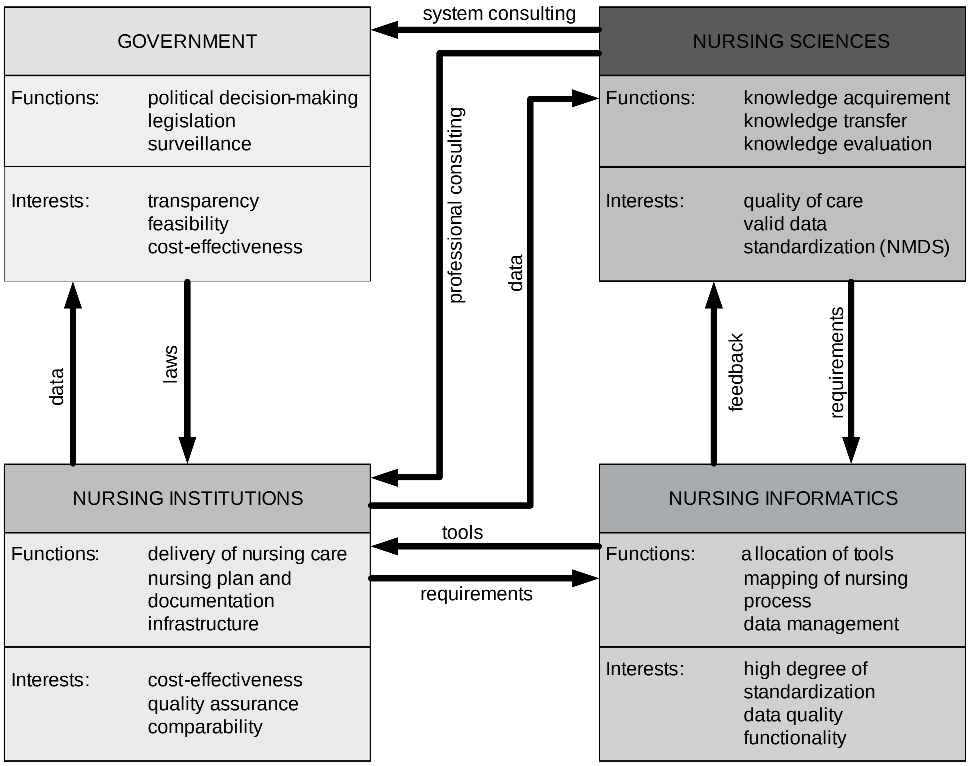
The government functions in association with the nursing system and nursing documentation can be summarized in the tasks of political decision-making, legislation for the nursing system and control function for the adherence of legal regulations within the framework of nursing people in need of care. These functions under the aspects of transparency, feasibility, and cost-effectiveness are converted into state parameters and initiatives for the achievement of these parameters.

Nursing sciences tend to acquire perception and transfer of knowledge under the aspects of nursing outcome quality and process standardization. Nursing sciences define the parameters and conditions of measuring nursing performance.

FUTURE TRENDS

The acquisition of knowledge, transfer of knowledge, and evaluation of nursing knowledge in the sense of an increase in quality within nursing as basic functions can be accredited to the nursing sciences as representatives of theoretical nursing. In this context, they have a great interest in the completeness, comparability and interoperability of the documented nursing data in order to fulfill their function. It can be assumed that the

Figure 1. Interaction between government, nursing sciences, nursing institutions and nursing informatics



question of terminological standardization as a basis for a uniform nursing language will be more strongly internationalized in the future. This means that national factors in this relation will move into the background in order to achieve international comparability of nursing. In return, this will have a direct effect on the development of NMDS which will be internationally useable and transparent and free of national stamping.

The institutions where nursing takes place are responsible for providing the necessary infrastructure and direct nursing process, that is, diagnoses, planning, implementation, documentation, and evaluation of the nursing provided. The delivery of nursing care is subject to the criteria of efficiency, quality, and individualization of nursing for the clients. The comparability of nursing is only secondary for evidence-based nursing here. But evidence-based nursing in particular will flow more strongly into all elements of the nursing process in the future, both on a national as well as international level. Here, more intensive integration of applied nursing with nursing science will be demanded in order to overcome the Theory-Practice-Gap as described in literature. An essential component for bridging this gap will be nursing informatic science.

The government as political constructor will have an increased interest in the international comparability of its nursing system when the objectives of transparency and political margin of national nursing systems are achieved and when it is in the position to establish an overall picture regarding efficiency, quality and cost of the national nursing system on the basis of the data obtained from NMDS.

CONCLUSION

Nursing sciences the world over are shaped by a trend of standardization. Here the focus is on the development of a uniform nursing terminology and a standardized definition of the nursing process. An essential basis for achieving this objective is the digitalization of the nursing process on the one hand and the development of Nursing Minimum Data Sets (NMDS) on the other. At the same time one can observe that health systems and therefore also the nursing systems are subject to increased cost pressure, which in turn leads the financiers and legislators to claim not only increased transparency of structures, but also to perform their mandate of configuring these systems on the basis of comparable and valid data from nursing care deliverers. Against this background, the state institutions also have a distinct interest in the development of NMDS. The motivation situation between the nursing sciences and practical nursing as well as that of the government is not consistent. Whilst the government needs consolidated data as a foundation of configuration of the nursing system, the emphasis of the nursing sciences and practical nursing is directed towards supporting the nursing process and the scientific acquisition of knowledge. On an international level, projects for the development of an NMDS were carried out in different countries, whereby the objectives of the individual projects more strongly support governmental or nursing motivation. The result, however, of all these projects is that, in the end, both objectives have to be fulfilled to a certain degree in order to be able to implement NMDS successfully. It can be assumed that this phase of considering national NMDS will continue in the next few years. This is because the advances in the respective countries are indeed different, but the question of developing an international NMDS will increase in importance in the medium-term. This is all the more so because the nursing sciences show clear international focus within the framework of standardization even today, and because international comparability of nursing systems is advantageous from a governmental point of view. Even if there is great governmental influence on the forming of the nursing system and, therefore, on the development of NMDS through its capacity as legislator and financier, it is nevertheless to be assumed that the interaction between government and nursing sciences or applied nursing resembles that of communicating vessels. This is because the nursing sciences deliver the necessary standardization and scientific penetration of the system which, in turn, represents the foundation for political decision-making within the framework of the formation of the nursing system.

REFERENCES

Australian Institute of Health and Welfare. (2006). Nursing and midwifery labour force 2004. National health labour force series no. 37. Cat. No. HWL 38. Canberra, Australia: AIHW. Retrieved February 9, 2008, from http://www.aihw.gov.au/publications/index.cfm/title/10380

Berthou, A., Junger, A., & Kossaibati, S. (2005). Test des Schweizer nursing minimum data sets (CH-NMDS) (Test II) Schlussbericht; (E: Test of the Swiss Nursing Minimum Data Sets (CH-NMDS) (Test II). Retrieved February 9, 2008, from http://www.isesuisse.ch/nursingdata

Blegen, M., Goode, C., & Reed, L. (1998). Nurse staffing and outcomes. *Nursing Research, 47,* 43-50.

Czaplinski, C., & Diers, D. (1998). The effect of staff nursing on length of stay and mortality. *Medical Care, 36*, 1626-1638.

Fawcett, J., & Russell, G. (2001). A conceptial model for nursing and health policy. *Policy, Politics & Nursing Practice, 2,* 108-116.

Goossen, W.T.F. (2001). Exploiting the nursing minimum data set for the Netherlands. *Medinfo,* 1334-1338. IMIA 2001. Retrieved November 4, 2006, from EBSCOhost Business Source Elite database.

Gorenberg, B.D., Alderman, M.C., & Cruise, M.J. (1991). Social policy statements: Guidelines for decisionmaking. *International Nursing Review, 38*(1), 11-13.

Hinz, M., & Dörre, F. (2002). Von der lokalen Patientenakte zum internationalen Vergleich – Modernes Datenmanagement mit der ICNP (E: From local patient recort towards international comparison – modern data management by means of ICNP). Pflegezeitschrift 4/2002; pp 289 – 292.

ICN International Council of Nurses (2003). *ICNP® Internationale Klassifikation für die Pflegepraxis*; Hinz M., Dörre F., König P., Tackenberg P.; deutschsprachige Printversion 2003 by Verlag Hans Huber, Bern. (E: International classification for nursing practise; german printet edition).

ICNP® International Classification for Nursing Practice. A project of the International Council of Nurses (ICN).

Johnson, M., Maas, M.L., & Moorhead, S. (Eds.). (2005). Pflegeergebnisklassifikation NOC; deutschsprachige Ausgabe übersetzt, bearbeitet und herausgegeben von Andreas Büscher und Peter Tackenberg;

Verlag Hans Huber, Hogrefe AG, Bern, 2005; Originalausgabe: *Nursing Outcomes Classification NOC Second Edition*. Mosby 2000; St. Louis, MO.

Lamb, G.S., Mowinski Jennings, B., Mitchell, P.H., & Lang, N.M. (2004). Quality agenda: Priorities for action. Recommendations of the American academy of nursing conference on health care quality. *Nursing Outlook, 52,* 60-65.

Maas, L.M., & Delaney, C. (2004). Nursing process outcome linkage research – issues, current status and health policy implications. *Medical Care, 42*(2), 40-45.

Marin, H., Rodrigues, R., Delaney, C., et al. (2000). *Building standard-based nursing information systems.* Washington, DC: Pan American Health Organisation/ World Health Organisation, Division of Health Systems and Service Development.

NANDA International. (2005). NANDA Pflegediagnosen, Definition und Klassifikation 2005-2006; Deutschsprachige Ausgabe herausgegeben von Jürgen Georg; Verlag Hans Huber, Hogrefe AG, Bern; 2005; Originalausgabe: „Nursing Diagnoses: Definitions & Classification 2005 – 2006“; *NANDA International*, Philadelphia, USA.

Nonn, C.R., Mayer, H., & Evers, G.C.M. (2002). Pflegerische Leistung auf Fachabteilungen, Stationen und für Patientengruppen – Ein Vergleich auf der Basis Minimaler Pflegedaten (Nursing performance in specialist departments, medical departments and patient groups – a comparison by minimum nursing data). *Pflege, 15*, 69-77. Verlag Hans Huber Bern 2002. Retrieved November 4, 2006, from EBSCOhost Business Source Elite database.

NURSING data. (2003). Project NURSING data in Switzerland. Retrieved February 8, 2008, from http://www.isesuisse.ch/nursingdata/, Modellbeschreibung/ Model: http://www.isesuisse.ch/nursingdata/de/dokumente/modele_systeme_information_d.pdf, Schlussbericht/Final report: http://www.isesuisse.ch/nursingdata/de/dokumente/test02_rapport_final_0602-d.pdf .

Nursing Interventions Callsisfication (NIC). (1996). *Iowa Intervention Project* (2nd ed.). St. Louis, MO: Mosby.

Oddi, L.F., & Cassidy, V.R. (1990). Nursing research in the United States: The protection of human subjects. *Int. J. Nurs. Stud., 27*(1), 21-34.

Schaubmayr, C. (2004). Bedeutung einer EDV-gestützten Pflegedokumentation für das Pflegemanagement – Analysen, Vorschläge und Vision; unveröffentlichte Dissertation (Relevance of an IT-based nursing documentation for nursing managment – analysis, recommendations and visions). UMIT; Hall i.T.

Stark, S., & Hölzer, S. (2005). Berthou A. Pflegeinformationssysteme: Das Projekt NURSING data in der Schweiz (Nursing informations systems: The Swiss Nursing data project). *Swiss Medical Informatics (SMI), 54*, 3-5.

Turtiainen, A.-M., Kinnunen, J., Sermeus, W., & Nyberg, T. (2000). The cross-cultural adaptation of the Belgium nursing minimum data set to Finnish nursing. *Journal of Nursing Management, 8,* 281-290. Blackwell Sience Ltd. Retrieved November 4, 2006, from EBSCOhost Business Source Elite database.

Volrathongchai, K., Delaney, C.W., & Phuphaibul, R. (2003). Nursing minimum data set development and implementation in Thailand. *Journal of Advanced Nursing, 43*(6), 588-549. Retrieved November 4, 2006, from EBSCOhost Business Source Elite database.

Werley, H.H., Devine, E.C., Zorn, C.R., Ryan, P., & Westra, B.L. (1991, April). The nursing minimum data set: Abstraction tool for standardized, comparable, essential data. *American Journal of Public Health, 81*(4), 421-426. Retrieved November 4, 2006, from EBSCOhost Business Source Elite database.

ADDITIONAL READING

Australian Institute of Health and Welfare Annual Report 2004-05. Retrieved February 9, 2008, from http://www.aihw.gov.au/publications/aus/ar04-05/ar04-05-c01.pdf; http://www.aihw.gov.au/publications/aus/ar04-05/ar04-05-c02.pdf

Axelsson, L., Björvell, C., Mattiasson, A.-C., & Randers, I. (2006). Swedish registered nurses' incentives to use nursing diagnoses in clinical practice. *Journal of Clinical Nursing, 15*, 936-945. Blackwell Publishing Ltd. Retrieved November 4, 2006, from EBSCOhost Business Source Elite database.

Bostick, J.E. (2004). Relationship of nursing personnel and nursing home care quality. *Journal of Nursing Care Quality, 19*(2), 130-136. Lippincott Williams & Wilkins, Inc. Retrieved November 4, 2006, from EBSCOhost Business Source Elite database.

Butler, M., Treacy, M., Scott, A., Hyde, A., MacNeela, P., Irving, K., et al. (2006). Towards a nursing minimum data set for Ireland: Making Irish nursing visible. *Journal Compilation 2006,* 364-375. Blackwell Publishing Ltd. Retrieved November 4, 2006, from EBSCOhost Business Source Elite database.

Coenen, A., Weis, D.M., Schank, M.J., & Matheus, R. Describing parish nurse practice using the nursing minimum data set. *Public Health Nursing, 16*(6), 412–416. Retrieved November 4, 2006, from EBSCOhost Business Source Elite database.

Fischer, W. *Wie kann die Pflege zeigen, was sie leistet? (How can nursing show its contribution?).* Retrieved February 8, 2008, from http://www.fischer-zim.ch/artikel/Pflege-Leistung-9606-SBK.htm

Goossen, W. (2006). Representing nursing information in EHR and message standards: Analyzing, modelling, implementing. In *European Nursing Informatics Conference*, Osnabrück, Germany. Retrieved November 4, 2006, from EBSCOhost Business Source Elite database.

Goossen, W.T.F., Epping, P.J.M.M., Van den Heuvel, W.J.A., Feuth, T., Frederiks, C.M.A., & Hasman, A. *Development of the nursing minimum data set for the Netherlands (NMDSN): Identification of categories and items*. Retrieved November 5, 2006, from EBSCOhost Business Source Elite database.

Griens, A.M.G.F., Goossen, W.T.F., & Van der Kloot, W.A. (2001). Exploring the nursing minimum data set for The Netherlands using multidimensional scaling techniques. *Journal of Advanced Nursing, 36*(1), 89-101. Blackwell Science Ltd. Retrieved November 4, 2006, from EBSCOhost Business Source Elite database.

Holzer, E., & Hauke, E. (Hrsg.) (2004). Gesundheitswesen – vom Heute ins Morgen (E-health care – from today into tomorrow). In Staudinger R. (Ed.), *Gesundheitswesen und die Rolle der Politik*.WUV Wiener Universitätsverlag 2004; pp 91 – 99.

Huang, Z.-B., Neufeld, R.R., Lkourezos, A., Breuer, B., Khaski, A., Milano, E., et al. (2003). Sociodemographic and health characteristics of older Chinese on admission to a nursing home: A cross-racial/ethnic study. *Journal of the American Geriatrics Society, 51,* 404-409. Retrieved November 4, 2006, from EBSCOhost Business Source Elite database.

Jensdóttir, A.B., Rantz, M., Hjaltadóttir, I., Gudmundsdóttir, H., Rook, M., & Grando, V. (2003). International comparison of quality indicators in United States, Iceland and Canadian nursing facilities. *International Nursing Review, 50*, 79-84. International Council of Nurses 2003; Retrieved November 4, 2006, by EBSCOhost, Bussiness Source Elite database.

Keenan, G., Stocker, J., Barkauskas, V., Treder, M., & Heath, C. (2003, July/September). Toward collecting a standardized nursing data set across the continuum: Case of adult care nurse practitioner setting. *Outcomes Management, 7*(3), 113-120. Wolters Kluwer Health/Lippincott Williams & Wilkins, Inc. Retrieved November 15, 2006, by OVID.

Mitty, E.L. (2003, August). Policy perspectives: Assisted living and the role of nursing: As many as half a million people reside in assisted living facilities, the regulations of which vary from state to state. Nurses have an opportunity – and an obligation – to help develop policies. *American Journal of Nursing, 103*(8), 32-43. Retrieved November 15, 2006, by OVID.

Nursing Diagnosis Definitions & Classification 1999 – 2000 (1999). *North American Nursing Diagnosis Association*. Philadelphia, PA: NANDA.

Ryan, B. (1998, July 14). Rural medicine: What role should nurse practitioners play? *Canadian Medical Association Journal, 08203946, 159*(1). Retrieved October 23, 2006, from EBSCOhost Business Source Elite database.

Ryan, T., Holmes, B., & Gibson, D. (1999). *An national minimum data set for home and community care*. AIHW cat. No. AGE 13. Canberra, Australia: Australian Institute of Health and Welfare (Aged Care Series). Retrieved February 9, 2008, from http://www.aihw.gov.au/publications/index.cfm/title/4600

Staudinger, B. (2006). Voraussetzung und Funktionalitäten von Pflegedokumentations- und Informationssytemen in Österreichischen Alten- und Pflegeheimen (Proconditions and functionalities of nursing documentation and information systems in Austrian elderly care units). Unveröffentlichte Magisterarbeit; *Hall i.T.*, p. 14-15.

Wieteck, P. (2004). Standardisierte Fachsprache zur Pflegeprozessdokumentation: Zwischen Einheit und Vielfalt; Pflegezeitschrift (Standardized terminology in nursing process documentation: Between uniformity and variaty). In *Fachzeitschrift für stationäre und ambulante Pflege*; Verlag W. Kohlhammer, Stuttgart, GE, 3/2004, p. 193-196.

Yegdich, T., & Cushing, A. (1998). An historical perspective on clinical supervision in nursing. *Australian and New Zealand Journal of Mental Health Nursing, 7,* 3-24. Retrieved November 4, 2006, from EBSCOhost Business Source Elite database.

KEY TERMS

Documentation System: An IT-Tool to register nursing action; nevertheless, this system can still be paper based.

Nursing Classification: Or nursing classification system. A catalogue structuring nursing action. The focus is on nursing diagnosis (NANDA), nursing intervention (NIC), or nursing outcome (NOC). All classification systems intend to be a basis for possible quantification and also comparability of nursing performance.

Nursing Documentation: Contains more than the registration of delivered nursing action. A complete nursing documentation also comprises the assessment, plan, and evaluation—altogether representing the nursing process.

Nursing Minimum Data Set (NMDS): Dependent on the regional scientific background, NMDS is used for the concept of a data structure or a list of nursing interventions being documented in a nursing documentation system.

Nursing Process: A concept of how nursing care should be provided from a nurses' point of view. The nursing process was internationally published by WHO in the 1980s and originally consisted of four steps (assess, plan, intervene, evaluate). Due to its importance, documentation was additionally introduced to the nursing process.

Nursing Terminology: A nomenclature which enables the comparability of nursing performances.

A Role-Based Agent-Oriented Approach to Medical Device Integration

Francesco De Mola
Università di Modena e Reggio Emilia, Italy

Giacomo Cabri
Università di Modena e Reggio Emilia, Italy

BACKGROUND

Monitoring heterogeneous medical devices is a key issue for efficient medical information management in high-quality health care delivery. Collecting vital signs and integrating data from various sources is very important in order to make precise diagnosis and provide effective emergency response. Current commercial rules and the scarce adoption of standard protocols are the first main obstacle to achieve a real integration.

We propose a work-in-progress, role-based, agent-oriented solution to overcome this obstacle: proper mobile agents are in charge of acquiring data from each kind of medical device (e.g., electrocardiograms, blood pressure, and oxygen saturation gauges), providing a uniform interface toward the business logic; in its turn, each medical device is provided by a vendor-specific role, embodying the capability to interface with the device by means of its own protocol. Thus, each time a new medical device requires the integration in a wider system, a generic mobile agent migrates to the device node; here, it assumes the role of the specific device manufacturer and starts delivering the acquired data hiding the heterogeneity of the device.

We will propose such a solution describing the implementation of a demonstrative prototype exploiting the RoleSystem infrastructure, which dynamically makes role capabilities available to the agents.

INTRODUCTION

Information technology can provide an interesting support to manage applications in health care scenarios. The tasks involved in health care exhibit wide heterogeneity and are often strictly related to one another. This demands the adoption of innovative and powerful paradigms, in particular to achieve device integration that is more dynamic and flexible than those usually adopted for traditional distributed systems.

Agents are software components that exhibit a high degree of autonomy (Jennings, 2001) and provide a useful paradigm to meet health care requirements in general (Nealon & Moreno); in the literature, we also find several approaches for medical applications (Lhotska & Stepankova, 2005; O'Donoghue & Herbert, 2006; Shnayder, Chen, Lorincz, Fulford-Jones & Welsh, 2005). This chapter aims to show a possible agent-oriented solution for the easy integration of medical devices facing heterogeneity by means of the concept of agent role (Cabri, De Mola, Quitadamo & Zambonelli, 2006). A first prototype, UbiMedic-RED (Role-Enabled Devices) has been developed on the basis of the proposed approach. In particular, the implemented system is based on the *UbiMedic framework* (De Mola, Cabri, Muratori, Quitadamo & Zambonelli, 2006), a context-aware middleware built on top of the multi-agent platform JADE (JADE, 2001). It provides several services of high and low (core services) levels to support medical and territorial emergencies in distributed contexts. The concept of role has been introduced in UbiMedic, thanks to the integration with *RoleSystem* (Cabri, Leonardi & Zambonelli, 2003b), an interaction infrastructure implementation allowing agents to dynamically assume roles and interact accordingly. Section Staring Points presents the starting points of our work: the UbiMedic framework and the Rolesystem infrastructure. The rest of the article proposes the UbiMedic-RED architecture based on the previously introduced systems, exploiting an application scenario.

STARTING POINTS

UbiMedic aims at building a distributed context-aware platform with mobile agents implementing helpful services for user applications in medical domain, focused on real-time monitoring and delivering vital signs and medical information. UbiMedic is an agent-based framework, designed to take into consideration the importance of portability even on small devices (e.g., PDAs or other limited user's terminals) (De Mola et al., 2006). The framework exploits the execution environment and the facilities provided by the JADE agent distributed platform. The latter is composed of several *Peripheral Containers*, representing different nodes with their own JVM, and a centralized *Main Container*, providing the basic facilities for the platform management. In particular, the *Directory Facilitator (DF)* provides a yellow page service useful for agent discovery and registration.

About monitoring and integrating physically distributed medical devices, the framework proposes a three-level decomposed approach related to three corresponding agents.

The main idea of each application accessing a medical device (e.g., an electrocardiograph or pulse oximeter sensor) is the definition of the following three kinds of agents (Figure 1): a *User Interface Agent* (UIA) in charge of managing user interactions by means of a more or less complex GUI providing all the suitable tools to end-user monitoring of the detected patient data; a *Physical Resource Interface Agent* (PRIA), which has to collect and make available the patient's vital signs measured by medical devices to requesters all over the platform; and a *Proxy Agent* (PA), an intermediate entity that includes the proper logic to process and filter the data collected by PRIA according to the specific medical device typology and to the particular goals of the application. The PA is also the logic unit where further intelligent and cooperative diagnostic capabilities can be included; for example, exploiting the agent social ability, the PA can compare the received data with medical traces coming from other devices or databases before returning to the UIA the processed diagnostic results.

Rolesystem is an interaction infrastructure (Cabri et al., 2003b) that implements the role interaction model of BRAIN (Cabri, Leonardi & Zambonelli, 2003a). It is completely written in Java and can be associated with the main agent platforms; in the following we consider JADE (2001) as the agent platform.

As shown in Figure 2, the Rolesystem infrastructure is divided into two parts: the upper one, which is independent of the agent platform; and the lower part, which is bound to the chosen agent platform. Our effort was in the direction of reducing the platform-dependent part.

In applications exploiting the Rolesystem infrastructure, agents can be seen by two points of view: they are both *subjects* of the role system and *agents* of the JADE platform. Accordingly, an agent is composed of two layers: the *Subject Layer*, representing the subject of the role system independent of the platform; and the *Wrapper Layer*, which is the JADE agent in charge of supporting the subject layer. For each context/environment, a specific agent called *Server Agent* is in charge of managing the roles and their interactions. It interacts with the wrapper layer of agents by exchanging ACL messages formatted in an appropriate way.

In the following sections, we will show how the UbiMedic framework has been combined with the RoleSystem infrastructure.

Figure 1. The three-level decomposition of UbiMedic for device monitoring

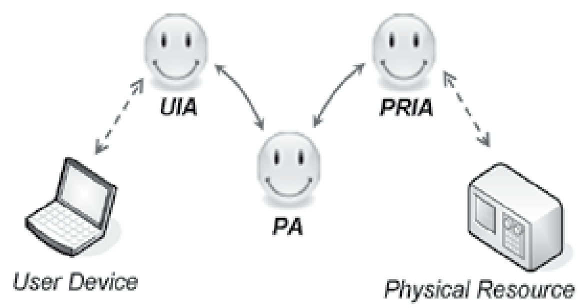
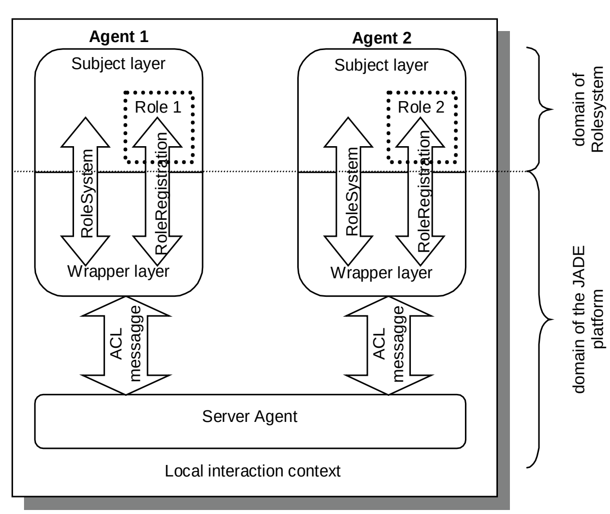


Figure 2. Domain separation in Rolesystem



CASE-STUDY SCENARIO

A simple example is exploited in the following sections to show the use of UbiMedic-RED. Let us consider a typical integration issue in medical device monitoring contexts: the interconnection of two heterogeneous devices provided by two vendors, allowing the end users (i.e., doctors and specialists) to remotely display on their terminals the vital signs measured by both the devices. In this scenario, *DeviceA* (provided by *VendorA*) is able to measure a patient's electrocardiogram (ECG) and pulse oximetry, whereas *DeviceB* (provided by *VendorB)* deals only with ECG. *DeviceA* includes also a thermometer provided by a third vendor (*VendorC*). In the developed prototype, these devices have their own graphical user interfaces for on-site display and control, their private internal algorithms for data generation, and their personalized output protocols and data structures. In particular, *DeviceA* fills in a proper file with the tuples of the collected data (ECG and oximetry) for each sampling instant, whereas *DeviceB* exposes a method returning the current detected ECG value.

The end user can identify and display the information acquired by the desired devices, obtaining a uniform visualization for the same data typology. Then, the possible differences among the data original formats coming from different sources are transparent to the end user.

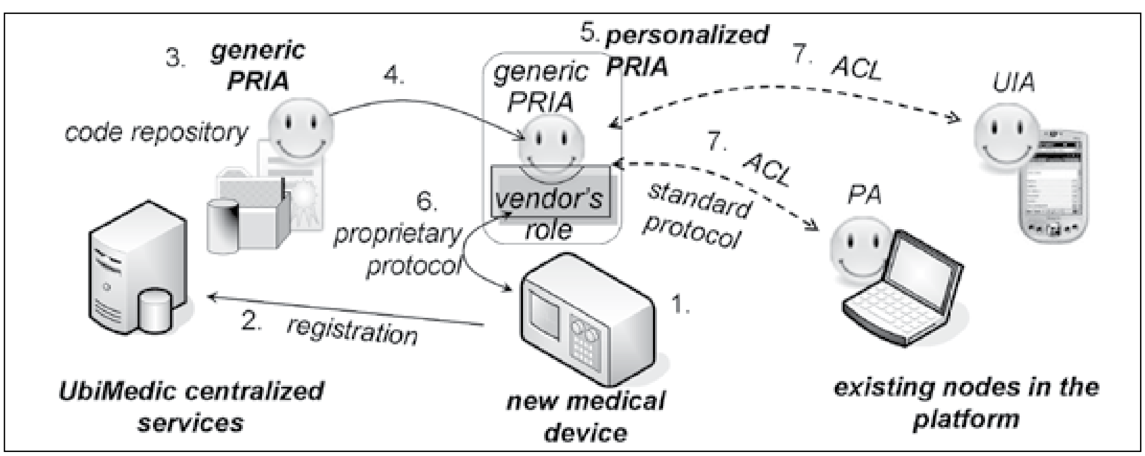
UbiMedic-RED ARCHITECTURE

UbiMedic-RED (Role-Enabled Devices) provides a role-based solution to *heterogeneous device integration* issues. The key point of the architecture is the encapsulation of specific medical device handling within software agents. This is reached through an interaction model based on the *role* concept. Among other advantages, roles can be exploited to achieve *separation of concerns* between the *business* logic and the *interaction* logic: the first is embodied in the agent, the latter in the role (Cabri et al., 2003b). When an agent assumes a role, it achieves the capabilities of a new behavior. Appropriate roles can be defined to support the interactions between agents and medical devices; the role can embed the specific logic to access the device information. In this sense, roles promote context-dependent interactions, where contexts are, in this case, the devices. The role implementer is in charge of designing the role adopting the UbiMedic standard interface for this kind of role. Internally, the code within the role will handle the interaction with the specific device, whereas the PRIA agent remains completely unaware of such inner details.

The adopted infrastructure in charge of providing role capabilities to the agent is the *BRAIN* framework (Cabri, Ferrari & Leonardi, 2004), and in particular, the RoleSystem infrastructure associated with the JADE agent platform.

Figure 3 clarifies the sequence of steps designed in UbiMedic-RED according to the role-based approach. A new medical device joining the platform must be

Figure 3. Interactions among the UbiMedic-RED components



equipped with a vendor-specific role (its implementation could be provided by the device's manufacturer), enclosing the know-how required to interact with the specific medical device (step 1 in Figure 3). First, the new device addresses a registration request to a centralized UbiMedic service to connect to the platform (2). The system reacts with the generation of a new generic PRIA in a centralized node, properly equipped with a code repository (3); then, the new PRIA moves back to the node of the requesting device (4), where it assumes the role of the specific device manufacturer, resulting in a complete personalized PRIA (5). The assumed role enables the device to interact with the other agents in the platform, although using its own procedures and proprietary protocols; the PRIA acquires data from the medical device via its proprietary protocol managed by the embodied role (6), and at the same time delivers the collected data to the target UIAs or PAs, interacting via standard ACL (7).

Such a role-based approach provides flexibility in device integration, since a brand new device can autonomously install itself in the system. In addition, the PRIA representing a multifunctional device (e.g., a device capable of performing electrocardiogram, blood pressure measurements, etc.) can assume different roles for each functionality.

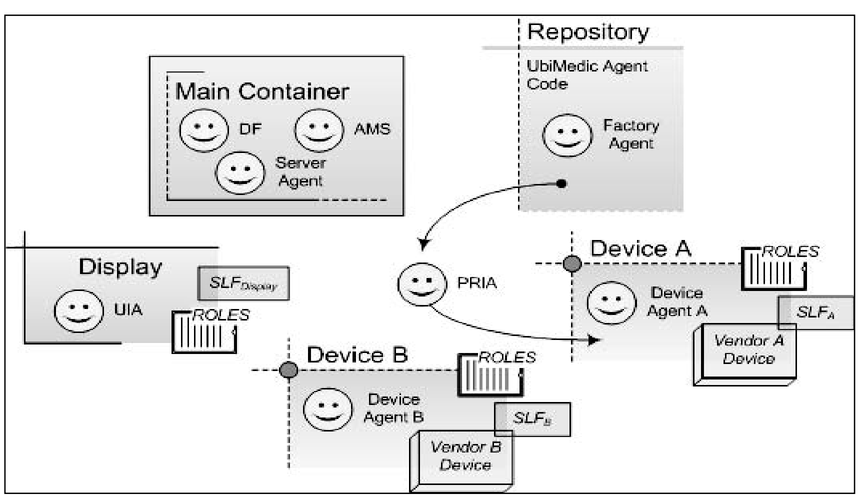
System Deployment

The nodes of the UbiMedic-RED platform are represented by common *JADE Containers*. The core of the system is the *Main Container*, including some UbiMedic centralized facilities and the RoleSystem *Server Agent* (besides to the standard JADE administration services) (Figure 4). Another *Container* called *Repository* implements the code repository in Figure 3; it is equipped with the code of the agents (PRIA) that can be dynamically created in the platform. A proper agent, the *Factory Agent*, is responsible for the generation on demand of such agents. The presence of such a centralized repository facilitates the system upgrade, avoiding multiple code updates in several locations and allowing a better management of the possible code releases.

The medical devices are represented in the agent platform by other peripheral containers. For each of them, a *Device Agent* is the first contact entity in the startup node configuration and the handshake phase to join the platform. This node also includes the abilities of acquiring data from the real device and the role packages related to the available medical functionalities. A PRIA can be considered the effective representative of the physical resource in the agent world only after these three steps: the agreement for a PRIA generation, the migration of the agent to the device node, and the assumption by the agent of the vendor-specific role. The PRIA works as a middleware stub to connect the device to the agent platform; its role behaves as a nonintrusive external wrapper of the device.

On the other side, some peripheral containers are built in the user devices with a proper UIA in charge to guide the medical practitioner in displaying the output of the various devices. The UIA must assume suitable

Figure 4. System deployment



roles to interact with the chosen data source. For this reason, all the peripheral containers provide a particular module, the *Subject Layer Factory (SLF)*, in charge to dynamically instantiate the proper roles for UIAs and PRIAs, according to the typology of the medical device chosen by the user, for the former, and to the vendor-based functionalities enforced by the medical device for the latter.

CONCLUSION

In this chapter, we presented UbiMedic-RED, a combined approach based on *agents* and *roles* to support the integration of heterogeneous medical devices.

Several advantages can be identified in the adoption of the proposed architecture. First, *separation of concerns* is achieved between the algorithmic and the interaction issues in developing agent-based applications; in fact, interaction issues with the physical devices are clearly separated from algorithmic issues, concerning processing activities to filter and transcode data in a common format for the whole platform. Then, since each device is equipped with an appropriate role, UbiMedic-RED can *integrate unknown kinds of medical devices*, allowing a sort of plug-and-play installation. In fact, the code related to the proprietary device protocols is completely embodied within the role. Another useful point is also the *reusability of solutions and experiences* allowed by the described agent and role-based model (e.g., the same code of an ECG PRIA can be used for electrocardiographs of various manufacturers). Besides, the presented solution results are suitable for a wide range of applications, not only medical device integration, such as m-commerce, warehouse management, robotics, and in general collaborative applications.

We are currently exploiting simulators of the medical devices, since role-enabled devices should be provided by the vendors along with appropriate (copyrighted) code. So the first step toward future works will concern the development of this code in collaboration with vendors and/or the exploitation of prototype devices whose specifications will be available for our purposes. Then, managing actual devices, we will be able to concretely evaluate the proposed integration approach. In connection with this, we point out that a performance analysis and discussion will be needed in order to evaluate the real applicability of the proposed solution. In fact, the health scenario requires not only appropriate abstractions and tools, but also quick responses, effectiveness, and robustness for the supporting infrastructures in order to save patients' lives.

REFERENCES

Cabri, G., De Mola, F., Quitadamo, R., & Zambonelli, F. (2006). Agent-based integration of medical devices for monitoring purposes. In *Proceedings of the Fourth Workshop on Agents Applied in Healthcare*, Riva del Garda, Italy

Cabri, G., Ferrari, L., & Leonardi, L. (2004). Agent roles in the BRAIN framework: Rethinking agent roles. *Proceedings of the IEEE Systems, Man and Cybernetics Conference*, The Hague, The Netherlands.

Cabri, G., Leonardi, L., & Zambonelli, F. (2003a). BRAIN: A framework for flexible role-based interactions in multiagent systems. *Proceedings of the 2003 Conference on Cooperative Information Systems (CoopIS)*, Catania, Italy.

Cabri, G., Leonardi, L., & Zambonelli, F. (2003b). Implementing role-based interactions for Internet agents. *Proceedings of the 2003 International Symposium on Applications and the Internet (SAINT)*, Orlando, Florida.

De Mola, F., Cabri, G., Muratori, N., Quitadamo, R., & Zambonelli, F. (2006). The UbiMedic framework to support medical emergencies by ubiquitous computing. *International Transactions on Systems Science and Applications, 1*(1), 1751–1461.

JADE. (2001). *Java agents development framework, TILAB, Torino*. Retrieved from http://jade.tilab.com

Jennings, N.R. (2001). An agent-based approach for building complex software systems. *Comm of the ACM, 44*(4), 35–41.

Lhotska, L., & Stepankova, O. (2005). Agent architecture for diagnostic and monitoring in medicine. *Proceedings of the 3rd Workshop of Agent Applied in Health Care*, Edinburgh.

Nealon, J.L., & Moreno, A. (2002). The application of agent technology to Health care. *Proceedings of the Workshop of Agent Applied in Health Care*, Lyon.

O'Donoghue, J., & Herbert, J. (2006). Profiled sensor data acquisition in a ubiquitous medical environment. *Proceeding of the Workshop on Ubiquitous and Pervasive Health Care*, Pisa, Italy.

Shnayder, V., Chen, B., Lorincz, K., Fulford-Jones, T.R.F., & Welsh, M. (2005). *Sensor networks for medical care*. Harvard University Technical Report TR-08-05.

KEY TERMS

Agent: A computational unit able to solve a problem and execute in open and dynamic contexts, thanks to the properties of reactivity, autonomy, proactivity, and sociality. The agents can run in a special agent platform such as JADE.

Device Integration: The connection of various devices in order to allow them to communicate, interact, and interoperate. The main issue to face to achieve integration is the heterogeneity of communication protocols and data formats

E-Health: The management of health-related issues by means of information and communication technology, exploited to improve quality and efficiency of health care services and facilitate caregivers in their work.

JADE: The most known and used agent platform. The framework, implemented in Java, provides many facilities to support the management of agent life-cycle and to develop distributed multi-agent systems.

Medical Device: A device able to detect and measure some kind of patient vital signs. The electronic devices usually provide outputs in their own format, based on the proprietary protocols and technical details implemented by the vendor, resulting in wide data heterogeneity and hard interoperability issues.

Role: A set of capabilities and expected behaviors related to the agent that plays such a role. The agent playing a role gets new functionalities and abilities to interact with the roles of the other agents.

Role-Enabled Device: A device equipped with a special role, which encloses its own communication details and is able to interact with the other agents of the platform by means of a standard interface. The role, possibly provided by the device manufacturer, acts like a sort of middleware stub to connect the device to the agent platform.

.

Security-Aware Service Specification for Healthcare Information Systems

Khaled M. Khan
Qatar University, Qatar

INTRODUCTION

With the rapid advancement of Web-based technologies, healthcare information systems are becoming increasingly heterogeneous in terms of their architecture, composition, and runtime characteristics. A healthcare system can be composed of several stand-alone service components, such as Web services available from various distributed sources for runtime execution. We use the terms *Web services* and *service* interchangeably in this chapter to refer to the same concept. A healthcare application system can be composed of multiple autonomous geographically dispersed software services. A healthcare software service is autonomous as it has its own executable code and uses its own data or files. The composition of a healthcare system can be dynamic or static, depending on how services are connected to each other to provide the services. Some of the services are downloaded directly from the Internet and executed dynamically with the application system. The use of independent services in the healthcare information system is appealing because it supports reusability of code and far efficient utilization of network resources, and it might be cost efficient.

Despite the benefits and usefulness of service computing, the characteristics of third-party services also present tremendous challenges for healthcare information security. As services are reused in far greater scale, there is a need for ensuring the security properties of third-party services to the composite healthcare system. Security is a systemwide property referring to the entire system composed of several services. Security architecture alone at the individual service level cannot protect a system unless the conformity of the security properties is checked with the enclosing application system in terms of required and ensured security properties. This issue of conformity of security properties between the individual service and the enclosing healthcare system has far-reaching implications because failure of this would seriously undermine the privacy of the patients and the reliability of the healthcare providers. Let us give a brief scenario (Han & Khan, 2006) in order to magnify the seriousness of the issue.

Consider a Web service that can generate various test data such as X-ray data, MRI images, CT scan images of the patients, and so forth. Another Web service uses these data as input and generates diagnosis reports to other Web services, which could ultimately produce prescriptions and prognoses of the patients' problems. Finally, the prescriptions are sent to another Web services that supplies medicines to the patients. In this scenario, we can see that the confidentiality and integrity issues of patient data are critical. Similarly, the authenticity of the service provider is an important issue, because the generated output of one service determines the outcome of the treatment produced by another service. Note that all these services are completely independent in terms of their development, their formation, and, most importantly, their security provisions. Different Web services have different security requirements as well as assurances during run-time processing of various sensitive patient data. The fundamental question is how service computing can cater various security requirements of different healthcare services in a composed federated e-healthcare system.

To address this question, this chapter proposes a framework of composing security properties for services used in healthcare information systems. It proposes a generic architecture of our proposed security-aware services and discusses the related issues. In contrast to most of the current initiatives that focus on incorporating existing security techniques into service computing, we see a great need and potential to develop a new framework for security-aware service composition in healthcare systems. The main objective of this chapter is to define a process that would enable us to compose various security properties of a federated healthcare information system that is assembled from various independent third-party software services.

BACKGROUND

The current practice of using software services in healthcare systems does not have any compositional process for properly evaluating the conformity of security properties between third-party services. This could dangerously lead to compromise to the enterprisewide security requirements of the healthcare systems, such as the confidentiality of patient data and the reliability of the services. Consequently, this risky practice of compositing systems of third-party services without due attention to the conformity of security properties could result in the degradation of system security. This practice virtually forces healthcare systems integrators to take undue risks by composing a system in order to achieve global services. Generally speaking, most systems integrators have neither the time nor the resources to examine the security properties of candidate software services. The required security functions that a healthcare system requires may not comply with the provided security profiles of a service.

A thorough examination of related literature reveals that very few research works in this area have been reported. Most research initiatives focus on how to make the individual services secure or how to make the enterprise systems more secure. In our opinion, these cannot solve the problem of "mutual" security conformity among various services

Web services paradigm is based on the Simple Object Access protocol (SOAP), the Web Services Description language (WSDL), and the Universal Description Discovery and Integration (UDDI). The Extensible Markup Language (XML) is the basic mechanism behind all of these technologies. Attempts have been made to extend WSDL to encode quality of services (QoS) properties (Curbera, Khalaf, Mukhi, Tai & Weerawarana, 2003) such as security, performance, and reliability. The aim was not to compose QoS properties along with the service composition, but rather to describe the QoS properties of Web services. None of the standards for XML-based definition language, such as Business Process Execution Language for Web Services (BPEL4WS) (Andrews et al., 2003), XML Process Definition Language (XPDL) (Kappel et al., 1995), Web Services Security Languages or WS-Security (Seely 2002), and Business Process Modeling Language (BPML) (BPMI, 2003), addresses the issue of the composition of security services to engineer the security properties of the composed Web services. WS-Security addresses Web service security by using existing security standards and specifications.

OUR FRAMEWORK

A healthcare software service may provide several functionalities; each may have quite different types of security properties from the others. For a software integrator, it is hard to predict what security properties a functionality supports unless they are well expressed with the interface of the functionality. Security properties are used for various reasons such as to authenticate a system, to authorize, and to ensure confidentiality and integrity. Examples of security properties can be passwords, private keys, secret keys, public keys, shared keys, and digital signatures. The protection of the service and its data and instructions is usually implemented with one or more security functions.

The security properties used to protect a functionality may have distinct security properties for a particular scenario. Security properties could also be grouped into two types based on their role in a functionality: security *precondition* and security *postcondition*. In order to verify whether the security precondition is met by the services user who wants to use the functionality, a reasoning engine is needed.

Based on these preliminaries, we can define the following types of *attributes* of a software service: (i) *the service has one or more functionality*; (ii) *a functionality may be protected with security functions*; (iii) *the security function is composed of security properties*; (iv) *the security properties can be classified as precondition or postcondition of the functionality*; and (v) *a reasoning engine is needed to verify the compliance of the security precondition and postcondition*.

Addressing the compositional security concerns discussed in the previous sections, we propose a framework for specifying these attributes at the architectural level of a service, as depicted in Figure 1.

The framework proposed in Figure 1 shows five distinct tasks. The task for identifying the functionalities of a service involves enlisting the associated functionalities of the service. For example, a *prepare diagnosis report service* may include functionalities such as *receiving pathological data*, *providing diagnosis report,* and so forth. In the next task, each of the identified functionalities is associated with its security properties. For example, *receiving pathological data*

Figure 1. A framework for specifying security attributes of the service

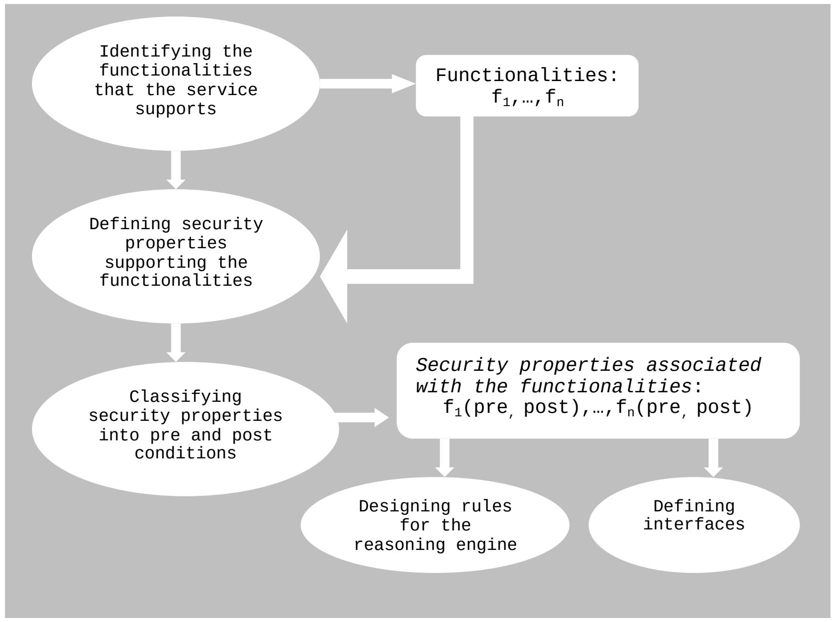
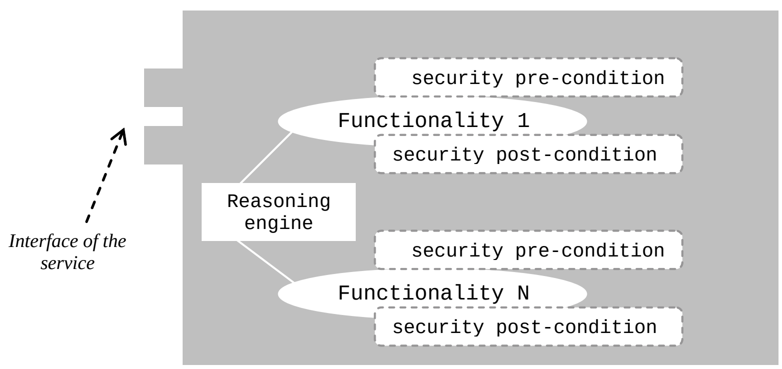


Figure 2. A healthcare service software with two functionalities



function may be supported with the encryption and digitally signed security functions. In task four, the employed security functions are classified into security precondition or postcondition of the function. For example, *providing diagnosis report* function may have encryption and digitally signed as its two security postconditions. The receiving pathological data function could have a security precondition such as encryption of the data. The specifying of security precondition and postcondition could be achieved by using the security characterization approach proposed in Khan and Han (2003). The designing rules for the reasoning engine function involves issues on how to check the conformity of the security precondition as well as postcondition. It could be a rule-based approach or any other approach. The final task defines the interfaces so that other applications or services may contact the service.

In our approach, the security properties are associated with the interface of the services in such a way that the other services or application systems could read the security properties along with the functionality of the service. The approach could also provide a mechanism so that the application could check the conformity between the application and services. The output of this framework could be a service with its functionalities mapped with the associated security precondition and postcondition, as shown in Figure 2.

The diagram depicts that a healthcare service provides two distinct functionalities to other services of applications. Each functionality has its own security precondition and postcondition. The service has one common interface to access the functionalities of the service. The employed security properties are mapped with the functionality that a service supports. In order to use any of these functionalities, other services or application systems must comply with the security precondition of the candidate functionality. In return, the functionality ensures the specified security postcondition.

CONCLUSION

The chapter has outlined the problems related to the issue of compositional security of healthcare software services at runtime. It has proposed a high-level framework to develop security-aware healthcare software services. A security-aware service has its functionality mapped with an associated security precondition as well as postcondition. In order to get the service of a functionality, other application systems or services are required to meet the security precondition of the functionality. In return, the healthcare service ensures security postcondition to the application systems. The chapter argues that different healthcare information systems may require from the services different types of security functions in order to protect the information. Our approach advocates a need-based, security-aware composition; that is, a healthcare application can only be composed with a service that could comply with its security requirements. Similarly, a service may also require certain security properties to be satisfied by the enclosing healthcare system in order to receive the service. We acknowledge that further investigations need to be done in order to build a complete framework. The framework also needs to be tested in a real-world situation.

REFERENCES

Andrews, T., et al. (2003). Business process execution language for Web services. Specification Version 1.1, IBM.

BPMI. (2003). *Web services*. BPMI technical report.

Curbera, F., Khalaf, R., Mukhi, N., Tai, S., & Weerawarana. (2003). The next step in Web services. *Communications of the ACM, 46*(10).

Han, J., & Khan, K. (2006). Security oriented service composition and evolution. In *Proceedings of the IEEE Asia Pacific Conference on Software Engineering, IEEE Computer Society,* Los Alamitos, California, 171–178.

Kappel, G., Lang, P., Rausch-Scott, S., & Retschitzegger, W. (1995). Workflow management based on objects, rules, and roles. *IEEE Computer Society Data Engineering Bulletin*, *18*(1), 11-18.

Khan, K., & Han, J. (2003). A security characterisation framework for trustworthy component based software systems. In *Proceedings of the IEEE Computer Software and Application Conference*, *IEEE Computer Society*, Los Alamitos, California, 164–169.

Seely, S. (2002). *Understanding WS-Security*. Technical report, Microsoft.

KEY TERMS

Healthcare Service Functionality: A functionality provided by a service is a task offered by the service to its environment. A healthcare service may have one or more functionalities to offer. The client healthcare application system could use these functionalities at runtime.

Healthcare Services: A platform-independent, self-contained software with defined functionality that can provide healthcare related computing services. It provides a standard way to integrate mechanism with healthcare applications over the Web. A service can perform one or more functionalities for the complex healthcare application system.

Security-Aware Service: A service that has its own security requirements to be satisfied by the service user; in return, it guarantees certain security requirements. The approach also provides a mechanism to verify the compliance of the security requirements of the service as well as the service user.

Security Function: The implementation of a security policy as well as a security objective. It enforces the security policy and provides required capabilities. Security functions are defined to withstand certain security threats, vulnerabilities, and risks. A security function usually consists of one or more principals, resources, security properties, and security operations.

Security Postcondition: An ensured security property is a postcondition in a sense that it is the responsibility of the service or the application system to maintain the committed security assurances during the composition.

Security Precondition: An invariant in a sense that other entities must satisfy this before a composition takes place. It is a precondition the service user must ensure to the service provider that the required security properties are met and their validity is ensured before a service can be obtained.

Security Property: An implementation element used in a security function. A set of security properties can form a security function. A security property is an element at the lowest level of the implementation.

Security Reasoning Engine: Verifies the security compliance between a service and the service user. The reasoning engine could use rules that comprise a set of criteria that could be used to assess security properties between two services. A security reasoning engine tests whether a security function has the desired security properties.

Similarity Metrics for Medical Image Registration

Roger Tait
Nottingham Trent University, UK

Gerald Schaefer
Aston University, UK

INTRODUCTION

In general, image registration algorithms can be classified as either landmark or intensity-based. Landmark-based registration consists of four main stages:

- During the feature detection stage, distinguishing characteristics such as corners, edges, centres of gravity, and so forth are identified, either manually or automatically. This identification of landmarks is performed on both reference (fixed) and sensed (moving) images.
- The optimisation stage (Jenkinson & Smith, 2001) controls estimation of transform parameters that geometrically map landmarks between fixed and moving image.
- Upon selection of appropriate transform parameters, pixel values which are mapped into noninteger coordinates are interpolated in order to establish their value. This represents the image resampling stage (Grevera & Udupa, 1998).
- The feature matching stage is achieved through the use of a similarity metric in which a degree of closeness or accuracy of alignment between corresponding landmarks is calculated.

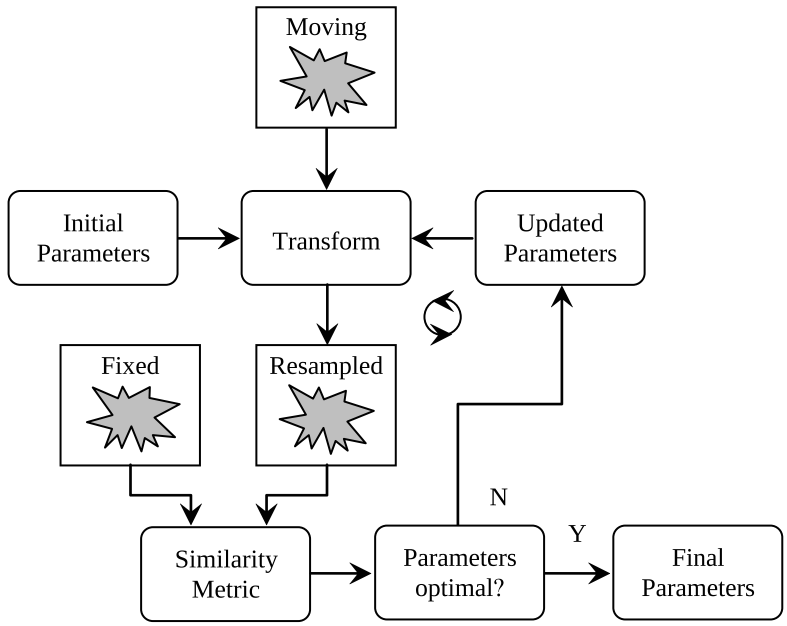
In intensity-based image registration methods, the feature detection stage is omitted. As a consequence, the transform optimisation and feature matching stages are performed using pixel intensities (or functions thereof) instead of landmarks. Intensity-based image registration algorithms comprise the following components:

- The spatial mapping of intensities throughout the alignment process is achieved with a transform component.
- An interpolation component is used to evaluate intensities at nondiscrete locations.
- The metric component calculates a measure of alignment accuracy.
- Optimisation of the similarity measure using a search space defined by transform parameters is achieved with an optimisation component.

The most important component of an image registration algorithm is the similarity metric used to determine when images are in accurate alignment (Penney, Weese, Little, Desmedt, Hill, & Hawkes, 1998). In Figure 1, the inputs to and output from a basic metric are illustrated. In general, a metric works by examining corresponding pixel values in both fixed and moving images and then formulating a measure of similarity based on the relationship between these intensities. The metric assumes that the relationship changes with variations in the spatial transformation used to map between images and a maximum similarity is achieved when the images are in close alignment (Brown, 1992).

Intensity equality which is high when pixels are similar is one such relationship employed as a similarity metric in single-modal registration where images are captured using the same sensor type. Total equality, however, is seldom reached due to noise and image acquisition inconsistencies. Additional robustness is therefore achieved by assessing the ratio of intensities and minimising the variance of such ratios. When images are acquired with different sensor types, as is typically the case in multimodal registration, an extension of the ratio method which maximises the weighted sum of variances can be employed. Alternatively, a relationship estimating the entropy of corresponding intensity pairs can be formulated where entropy, derived from information theory (Shannon, 1948), is the measure of the amount of information contained within a signal. Although many algorithms have been proposed, similarity calculation remains a complex task.

Figure 1. Flow diagram of similarity metric calculation



BACKGROUND

By comparing intensities, the metric component quantitatively measures how accurately fixed and moving images are aligned. The selection of a metric component is largely dependent on the type of registration problem to be solved (Roche, Malandain, Ayache, & Prima, 1999). For example, some metrics possess capture ranges that are well suited to images misaligned by a large transform. Other metrics, in contrast, are less computationally intensive but require initial transform parameters to be close to the optimum. During the alignment process, most metrics samples intensities over an entire image. Some metrics, however, employ a subset of samples drawn from the image. In both cases, similarity is calculated using intensities which fall within the boundary of the moving image. Intensity correlation has been used as a metric where a maximum similarity between fixed and moving images is searched for (Pratt, 1974). Using such an approach, high levels of alignment accuracy can be achieved by interpolating intensities before evaluating similarity. Such metrics are found predominantly in single-modal registration applications.

Although more difficult, the registration of images captured using different sensor types is commonplace in medical imaging applications. Viola (1995) suggests that for two images of differing modality, mutual information can be used as a measure of similarity where similarity is estimated using marginal and joint entropy based on probability distribution constructed using intensities from both fixed and moving images. As a consequence, mutual information can be described as the amount of information one image contains about another. Importantly, the ability to align multimodal images allows for the comparison of anatomical and functional data that can lead to a diagnosis which would be impossible to gain otherwise. For example, the evaluation of mutual information-based similarity metrics, used for the registration of brain scans, is presented by Holden et al. (2000).

SINGLE-MODAL REGISTRATION SIMILARITY METRICS

Intensity-based registration has been employed in the alignment of x-ray images and biomedical volume data (Russakoff, Rohlfing, & Maurer, 2003). In these applications, the cross-correlation of intensities can be used as a similarity metric. With such an approach, subpixel accuracy of alignment is achieved by interpolating intensities before similarity calculation. Traditionally, cross-correlation has been used to register translated images with only slight rotations and scalings. However, more sophisticated variants capable of handling complex geometric deformations are now commonplace. By employing phase correlation in the frequency, domain increased robustness to noise can be obtained (Castro & Morandi, 1987). Phase correlation involves the calculation of a cross-power spectrum for both fixed and moving images. The locations of peaks in both spectra are then identified and aligned.

The basic cross-correlation of intensities between fixed and moving images can be computed as:

$$CC(F,M)=\frac{\sum_{x=1}^{I}\sum_{y=1}^{J}F(x,y)M(x-u,y-v)}{\left[\sum_{x=1}^{I}\sum_{y=1}^{J}M^{2}(x-u,y-v)\right]^{\frac{1}{2}}} \quad (1)$$

where *F* and *M* are fixed and moving image intensity functions respectively. *I* and *J* are the number of pixel rows and columns. *x* and *y* are discrete pixel coordinates, while *u* and *v* represent the components of a transform. The measure of similarity produced by the metric forms a quantitative criterion which can be optimised in a search space spanned by the transform parameters. Importantly, the optimisation technique employed is used to produce derivatives of the similarity measure with respect to each transform parameter. The resulting derivatives are used to update the current transform and the process is repeated until an acceptable degree of alignment is achieved.

Although popular, correlation-based metrics are sensitive to the presence of outliers. These are artefacts which can appear in an image and cause biased registration. Using statistics, estimates of mean and covariance can be employed to increase robustness of the alignment process (Zhang & Rangarajan, 2004). Robustness is sought by weighting intensities in such a way as to reduce the effects of outliers and in some cases remove them completely. In particular, based on the computed distances of intensities from the mean, new intensities and a new mean are determined. Other common statistical approaches to increased robustness include weighted square error and nonquadratic error functions. An example of intensity-based registration that uses a robust correlation coefficient as a similarity metric was presented by Jeongtae and Fessler (2004). To evaluate the statistical properties of their approach, 2D-to-2D and 2D-to-3D registration of torso phantoms was performed. In each case, experimental results confirm an improvement in robustness to outliers. Unfortunately, correlation-based metrics are limited to the alignment of images from the same modality.

MULTIMODAL REGISTRATION SIMILARITY METRICS

In multimodal image registration applications, the data to be registered stem from two different capture devices, as opposed to single-modal tasks where images are retrieved using the same sensor type. According to Woods, Mazziotta, and Cherry (1993), the measure of alignment between images of differing modality can be based on the assumption that although different in value, regions of similar intensity in the fixed image will correspond to regions of similar intensity in the moving image. Furthermore, for pixels in corresponding regions, the ratio of their intensities should vary only slightly. Consequently, alignment can be achieved by minimising the average variance of this ratio. This can be realised through the construction of the joint probability distribution which represents a two-dimensional plot that contains combinations of intensities, taken from corresponding coordinates in both images (Hill, Hawkes, Harrison, & Ru, 1993). Hence, instead of identifying regions of similar intensity directly within the images, combinations thereof are analysed using the joint probability distribution.

Given fixed image *F*, moving image *M*, and transform *T*, the estimation of similarity using mutual information, as described by Mattes, Haynor, Vesselle, Lewellen, and Eubank (2001), is defined as:

$$S(F,M,T)=\sum_{l=1}^{MB}\sum_{k=1}^{FB}jpd(l,k)\log\frac{jpd(l,k)}{mmpd(l)\,fmpd(k)} \quad (2)$$

where *jpd* is a joint probability distribution extracted from both images using a set of samples and approximated as a histogram of intensities. Marginal probability distributions, *fmpd* and *mmpd*, are extracted from fixed and moving images respectively while *FB* and *MB* are the number of fixed and moving histogram bins. Each entry in the joint probability distribution denotes the number of times an intensity in the fixed image coincides with intensities in the moving image.

The marginal probability distributions, in contrast, are found by summing the number of times intensities appear in their respective images. In general, it is common for contributions to a probability distribution to be smoothed using a probability density function, for example, using Parzen windows. Once generated, both joint and marginal probability distributions are normalised and an estimation of similarity calculated. In this form, mutual information is a measure of the distance between the joint and marginal distribution of intensities. Mutual information is therefore a measure of dependence which is maximal when the two volumes are accurately aligned.

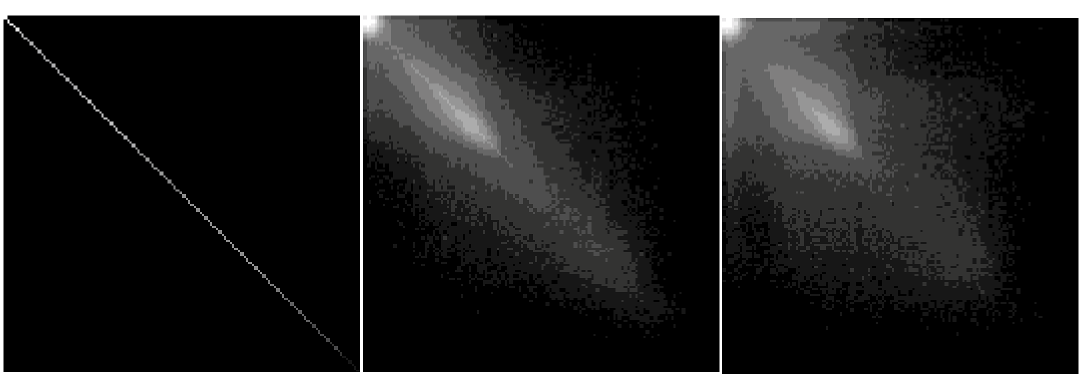
During the registration process, a variation in alignment between images causes changes in appearance of the joint probability distribution. When correctly aligned, corresponding structures in both images overlap causing a clustering of intensities combinations. Misalignment, in contrast, causes structures in the fixed image to overlap with different structures in the moving image which in turn results in the dispersal of intensity combinations within the joint probability distribution (Pluim, Maintz, & Viergever, 2003), as illustrated in Figure 2. Based on the changing regions within the joint probability distribution, measures of dispersion which guide the registration process have been proposed and successfully implemented (Maes, Collignon, Vandermeulen, Marchal, & Suetens, 1997). The concept of mutual information is a well recognised and accepted method of similarity calculation. The main advantage of employing mutual information is that the type of dependency between two variables does not have to be specified and as a result complex mappings can be modelled.

FUTURE TRENDS

To reduce the computational burden of similarity calculation, multiresolution schemes can be adopted. Subsampling at successive levels by means of interpolation is used to create a coarse-to-fine data structure (Thevenaz, Ruttimann, & Unser, 1998). Importantly, care needs to be taken to maintain geometric consistency at all resolution levels. Once subsampling is complete, similarity evaluation and transform parameter optimisation is performed. Crucially, optimisation proceeds through successive levels until the full resolution images are reached. As a consequence, the registration of large scale features is achieved first and only small corrections are required at progressively finer resolutions. Because the majority of iterations are performed at the coarsest level, where the fixed and moving images are small, a substantial saving in similarity evaluation burden is achieved.

The burden of similarity evaluation can also be alleviated using parallel processing (Tait & Schaefer, 2008). For example, a distributed blackboard architecture has been employed to provide an underlying worker/manager model in which image registration algorithms can be hosted (Tait, Schaefer, & Hopgood, 2006). Division of images into segments followed by allocation to multiple worker threads allow derivatives of the similarity measure with respect to each transform parameter to be estimated. The accumulation

Figure 2. Example joint probability distributions for cases where the moving image is obtained by rotating the fixed image by (from left to right) 0, 2, and 5 degrees. Intensity combinations disperse as the misalignment between images increases



and summation of local derivatives by the manager thread allows for the evaluation of similarity and the optimisation of transform parameters. Convergence of the optimisation process is then tested using the newly updated transform parameters. Depending on the success or failure of convergence testing, propagation of updated parameters and hence evaluation of a new transform occurs. Conveniently, the modular nature of the approach described permits different similarity calculation strategies to be employed without change to the underlying architecture.

CONCLUSION

In general, correlation-based metrics are best suited to images that do not contain high levels of prominent detail and where distinguishing information is contained within pixel intensities. These methods exploit directly the matching of image intensities without any structural analysis. However, as a consequence, they are sensitive to changes caused by noise and varying degrees of illumination. Traditionally, correlation has been used to register translated images. More sophisticated variants of the algorithm, however, which are capable of handling complex geometric deformations, are now commonplace. The main limitation of these metrics is that large translations and rotations are possible but render the algorithm obsolete due to the high computational burden associated with their evaluation. Although employed in alignment of images from the same modality, the development of mutual information as a similarity measure has successfully extended metrics to multimodal applications. As only few assumptions about the imaging process need to be made, mutual information-based similarity metrics can be generalised and applied to a wide variety of modalities. As similarity is calculated based on a subset of intensity samples, multimodal registration applications typically have reduced computational complexity.

REFERENCES

Brown, L.G. (1992). A survey of image registration techniques. *ACM Computing Surveys,* 325-376.

Castro, E.D., & Morandi, C. (1987). Registration of translated and rotated images using finite fourier transforms. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 9,* 700-703.

Grevera, G.J., & Udupa, J.K. (1998). An objective comparison of 3D image interpolation methods. *IEEE Transactions on Medical Imaging, 17,* 642-652.

Hill, D.L.G., Hawkes, D.J., Harrison, N.A., & Ru, C.F. (1993). A strategy for automated multi-modal image registration incorporating anatomical knowledge and imager characteristics. *Information Processing in Medical Imaging, 687,* 182-196.

Holden, M., Hill, D.L.G., Denton, E.R.E., Jarosz, J.M., Cox, T.C.S., Rohlfing, T. et al. (2000). Voxel similarity measures for 3-D serial MR brain image registration. *IEEE Transactions on Medical Imaging, 19,* 94-102.

Jenkinson, M., & Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis, 5,* 143-156.

Jeongtae, K., & Fessler, J.A. (2004). Intensity-based image registration using robust correlation coefficients. *IEEE Transactions on Medical Imaging, 23,* 1430-1444.

Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., & Suetens, P. (1997). Multi-modality image registration by maximization of mutual information. *IEEE Transactions on Medical Imaging, 16,* 187-198.

Mattes, D., Haynor, D.R., Vesselle, H., Lewellen, T.K., & Eubank, W. (2001). Non-rigid multi-modality image registration. *Medical Imaging 2001: Image Processing*, 1609-1620.

Penney, G.P., Weese, J., Little, J.A., Desmedt, P., Hill, D.L.G., & Hawkes, D.J. (1998). A comparison of similarity measures for use in 2D-3D medical image registration. *IEEE Transactions on Medical Imaging, 17,* 586-595.

Pluim, J.P., Maintz, J.B.A., & Viergever, M.A. (2003). Mutual information-based registration of medical images: A survey. *IEEE Transactions on Medical Imaging, 22,* 986-1004.

Pratt, W.K. (1974). Correlation techniques of image registration. *IEEE Transactions on Aerospace and Electronic Systems, 10,* 353-358.

Roche, A., Malandain, G., Ayache, N., & Prima, S. (1999). Towards better comprehension of similarity measures used in medical image registration. *Lecture Notes in Computer Science, 1679,* 555-566.

Russakoff, D.B., Rohlfing, T., & Maurer, C.R. (2003). Fast intensity-based 2D-3D image registration of clinical data using light fields. In *Proceedings of the 9th IEEE International Conference on Computer Vision* (pp. 416-423).

Shannon, C.E. (1948). A mathematical theory of communication (parts 1 and 2). *Bell System Technical Journal, 27,* 379-423 & 623-656.

Tait, R.J., & Schaefer, G. (2008, In press). Distributed medical image and volume registration. In *Encyclopedia of Healthcare Information Systems*. Hershey, PA: IGI-Global.

Tait, R.J., Schaefer, G., & Hopgood, A.A. (2006). Towards high performance image registration using intelligent agents. In *Proceedings of 13th International Conference on Systems, Signals and Image Processing* (pp. 435-438).

Thevenaz, P., Ruttimann, U.E., & Unser, M. (1998). A pyramid approach to subpixel registration based on intensity. *IEEE Transactions on Image Processing, 7,* 27-41.

Viola, P. (1995). *Alignment by maximization of mutual information*. Ph.D. Thesis. Artificial Intelligence Laboratory, MIT.

Woods, R.P., Mazziotta, J.C., & Cherry, S.R. (1993). MRI-PET registration with an automated algorithm. *Journal of Computer Assisted Tomography, 17,* 536-546.

Zhang, J., & Rangarajan, A. (2004). Affine image registration using a new information metric. *IEEE Transactions on Computer Vision and Pattern Recognition,* 848-855.

KEY TERMS

Image Registration: The process whereby two images, differing by a spatial transformation, are brought into geometric alignment.

Multimodal Registration: In multimodal registration applications, the reference and sensed images are captured using difference sensor types.

Reference (Fixed) Image: The term given to an original image or template. In visual inspection applications, the reference image represents an image containing a sample with an acceptable level of quality.

Sensed (Moving) Image: The term given to the image to be mapped to the reference image. In visual inspection applications, the sensed image represents an image of a sample for which the level of quality is to be determined.

Similarity Metric: A measure used to quantitatively judge how well a transformed moving image fits a corresponding fixed image.

Single-Modal Registration: In single-modal registration applications, the reference and sensed images are captured using the same sensor type.

Spatial Transformation: The transformation used to change the spatial relationship between pixels by mapping locations between fixed and moving images.

Source Localization of Subtopographic Brain Maps for Event Related Potentials (ERP)

Adil Deniz Duru
Bogazici University, Turkey

Ali Bayram
Bogazici University, Turkey

Tamer Demiralp
Istanbul University, Turkey

Ahmet Ademoglu
Bogazici University, Turkey

INTRODUCTION: DECOMPOSITION OF ERP DATA AND THE INVERSE PROBLEM

Event-related potentials (ERP) are transient brain responses to cognitive stimuli, and they consist of several stationary events whose temporal frequency content can be characterized in terms of oscillations or rhythms. Precise localization of electrical events in the brain, based on the ERP data recorded from the scalp, has been one of the main challenges of functional brain imaging. Several currentDensity estimation techniques for identifying the electrical sources generating the brain potentials are developed for the so-called neuroelectromagnetic inverse problem in the last three decades (Baillet, Mosher, & Leahy, 2001; Koles, 1998; Michela, Murraya, Lantza, Gonzaleza, Spinellib, & Grave de Peraltaa, 2004; Scherg & von Cramon, 1986). Generally, there are two main assumptions for the source models:

1. The number of sources are less than the number of measurements, which leads to the parametric methods that estimate the strength and position of the dipole sources by optimization methods (Mosher, Lewis, & Leahy, 1992).
2. The number of sources are more than the number of measurements that calls for the imaging methods searching for an optimum source distribution in a 3D brain image (Pascual-Marqui, Michel, & Lehmann, 1994). Since the number of source locations are much higher than the number of sensor measurements, the latter case is a severely underdetermined problem.

The uniqueness of the inverse solution for both cases can only be acheived by imposing additional constraints. Introducing a mathematical criterion such as the minimum norm, the maximum smoothness, or the minimum variance, or constraining the solution space is a common way to circumvent the uniqueness problem. Also, a well defined anatomical region where the orientation or position of the dipole is, a priori known can be used as an additional constraint. Other physiological constraints obtained from functional magnetic resonance imaging (fMRI), and Positron Emission Tomography (PET) studies may also be incorporated in the inverse solution methods. The sources underlying the ERP, which are rather distributed in the brain and nonstationary in time, also motivate the researchers to invent additional simplifications and analyses on the multichannel data before they apply source localizaton. Several studies aimed for isolating stationary frequency components in temporal windows, using time frequency analysis, and then applying source localization techniques to the scalp maps generated by these time frequency components. Source localization of epileptic discharges is performed after wavelet prefiltering to isolate the spiky waveforms from the background EEG (Ademoglu, Demiralp, Istefanopulos, Comu, & Baykan, 2004). Gonzalez Andino, Grave de Peralta Menendez, Lantz, Blank, Michel, and Landis (2001) proposed an eigen-based method to identify the location of each underlying source in the time-frequency plane that generates a certain topographic distribution. Unlike the previous time-frequency studies that use single channel data, Koenig, Marti-Lopez, and Valdes-Sosa (2001) perform time-frequency optimization on mutichannel data with the wavelet coefficients having minimum energy and maximum spatial smoothness distributions. A more general and unique space/frequency/time decomposition in terms of so-called atoms are proposed by Miwakaichi, Martinez-Montes, Valdes-Sosa, Nishiyama, Mizuhara, and Yamaguchia (2004), which overcomes the limitation of introducing artificial assumptions as indepence in

ICA or uncorrelatedness in PCA approaches (Zhukov, Weinstein, & Johnson, 2000).

BACKGROUND: SPATIAL DECOMPOSITION OF ERP MAPS

In all these studies, the problem of identifying the individul ERP components are treated by proposing a decomposition method that involves the time, frequency, channel, or their various combinations. Another common feature observed is that the concept of frequency is always associated with, and defined in, temporal domain. In Wang, Begleiter, and Porjesz (1998) the spatial ERP maps are enhanced by using first, a projection method that transforms the scalp poentials defined on a 3D surface to a 2D plane, and then applying a 2D multiresolution decomposition. This is the first approach that attempts to simplify the scalp topography by decomposing it into simpler maps using a spatial multiresolution.

A scalp topographic map for an ERP may be a superposition of several simpler subtopographic maps, each resulting from an individual electrical source located at a certain depth as a focal or extended activity. Furthermore, this source may have a temporal characteristics as an oscillation or a rhythm that extends in a certain time window, which has been a basis of assumption for the time-frequency analysis methods.

The subtopographic maps arising from sources located at certain depths with different extention of activities generally fall into different spatial frequency bands on the scalp. Therefore, a spatial frequency decomposition of a scalp potential distribution yields these subtopographic maps, whose source localizations will lead us to their electrical sources in the brain. It is essential to perform this spatial frequency decomposition on a realistic scalp surface without any distortion. To achieve this purpose, we propose a method that involves a realistic scalp model on which a 3D wavelet transform is performed, and the subtopographic maps obtained are source localized using inverse modeling.

FORWARD PROBLEM AND REALISTIC HEAD MODEL

Forward problem of EEG computes the electrical potentials on the scalp surface from the given source positions and strenghts in the brain. Boundary Element Method (BEM) is used with the Centre of Gravity (COG) approximation on a realistic head model (Hamalainen & Sarvas, 1989; Schlitt, Heller, Aaron, Best, & Ranken,1995). Realistic head model consists of brain, skull, and scalp surfaces. Surfaces of the brain, skull, and scalp are tesselated with 2000, 1200, and 1200 triangles respectively. If we have P different measurement sites in the conductor model, the forward problem can be formulated by the electrical potential $\mathbf{v}(\vec{s})$ at electrode location $\vec{s}$, due to a dipole at $\vec{r}$ with strength $\vec{m}$, as in Equation 1.

$$\mathbf{v}(\vec{s}) = \mathbf{H}(\vec{s},\vec{r})\,\mathbf{m}(\vec{r}) \quad (1)$$

Here, $\mathbf{v}(\vec{s})$ is $P \times 1$ electrical field vector, $\vec{m}$ is a 3×1 strength vector for a single current dipole source located at $\vec{r}$, and $\mathbf{H}(\vec{s},\vec{r})$ is a $P \times 3$ dimensional transfer function which depends on the dipole location $\vec{r}$, the measurement sites $\vec{s}$, and the geometrical and physical properties of the media.

BEM used for the forward problem is a numerical approximation technique which partitions the surface of a volume conductor into a triangular mesh. This technique has been used in dipole source localization of brain electromagnetic activity since the end of 80s. The human head is modeled as three homogeneous conductor layers: the outermost surface being the boundary for the scalp, the intermediate being for the skull, and the innermost being for the brain. The head model that we used in this study is developed using the average T1 weighted human brain MRI data provided by Montreal Neurology Institute (MNI). Statistical Parameter Mapping software 99 release (http://www.fil.ion.ucl.ac.uk/spm/), which is developed by Wellcome Institute is used for 3D segmentation of the brain, skull, and scalp. After segmentation, the surfaces are triangulated in order to generate the realistic head model. 30 channel electrode locations (Fp1, Fp2, F7, F3, Fz, F4, F8, FT7, FC3, FCz, FC4, FT8, T7, C3, Cz, C4, T8, TP7, CP3, CPz, CP4, TP8, P7, P3, Pz, P4, P8, O1, Oz, O2) are registered to the scalp surface by spline interpolation, using the T1 weighted MR data, the inion-nasion and pre-auricular coordinates, and the 10-20 Electrode Placement System. The surface of the scalp, which is represented with 16188 triangles, can be seen in Figure 1.

Figure 1. Realistic head model developed from T1 weighted human brain MRI data containing 16188 triangles and 7800 vertices

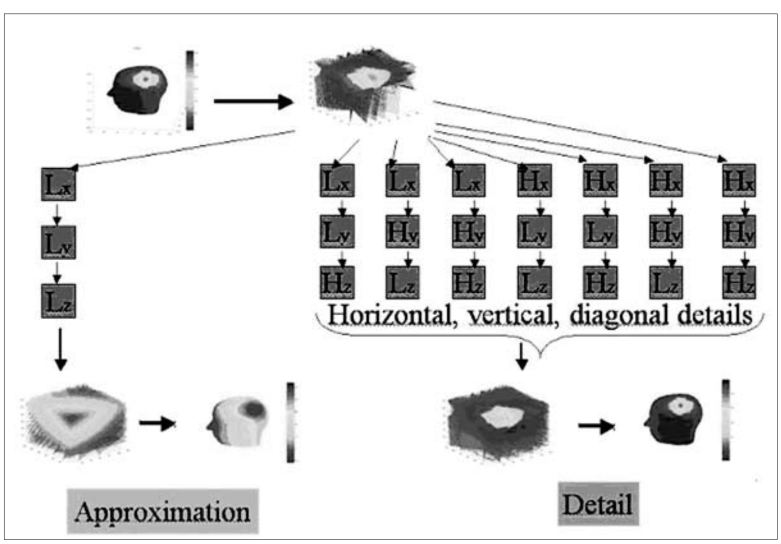
3D SPATIAL WAVELET DECOMPOSITION

Given a signal s of length N, the first step of Discrete Wavelet Transform (DWT) produces two sets of coefficients: approximation, and detail coefficients. These vectors are obtained by convolving the signal with the low-pass filter for approximation, and with the high-pass filter for detail, followed by dyadic decimation (downsampling) (Meyer, 1990). This decomposition splits data at level j into two components: the approximation at level j+1, and the detail at level j+1 (Daubechies, 1988; Grossman & Morlet, 1984; Mallat, 1989). For images, the algorithm is similar to the 1D case where the twoDimensional (2D) wavelets and scaling functions are obtained the by tensor product of respective 1D functions. 2D DWT leads to a decomposition at level j into four components: the approximation at level j+1, and the details in three orientations (horizontal, vertical, and diagonal). Similarly, to perform 3D wavelet decomposition, 1D DWT is applied on three directions on the 3D volume data, as seen in Figure 2. That leads to eight components from level j to j+1, one being the low pass filtered data, and seven others as the various combinations of low and high pass filtered data, in all three directions.

INVERSE PROBLEM

The inverse problem estimates the source positions, and their strength from multichannel ERP data. The MUSIC (Multiple Signal Classification) scanning algorithm is used for the inverse problem. MUSIC method is based on subdividing the brain tissue into a 3D grid, and computing the spatial power spectrum with an eigen-based approach for each voxel element. In order to do this, the transfer function $H(\vec{s}, \vec{r})$ in Equation 1 has to be computed numerically (Hamalainen & Sarvas,

Figure 2. 3D topography is converted to a 3D regular volumetric grid; by using 1D DWTs, at each stage, volumetric grid with raw data is decomposed into two sets called approximation and detail; finally, the original topograhies are regenerated from these volumetric maps



1989). In this study, the cortex of the brain is seleced as a solution space for MUSIC algorithm.

VALIDATION OF THE METHOD

In order to validate the method proposed, a simulated multichannel ERP data is generated by using COG approximation of the BEM. Two stationary radial point sources are assumed, one being superficial, while the other being deeper in the brain. Forward problem is solved for these point sources using BEM over the predefined head model given in Figure 1. The amplitude changes of the two point sources were sinusoidal with 128 data points, and they were temporally correlated. The inverse problem is solved using the total topographic activity. Although, temporally uncorrelated sources can be easily discriminated by the MUSIC algorithm, the source positions, in this case, are not clearly separated because of the temporal correlation of individual sources. MUSIC estimated the location of the superficial source from the total topography, which is a missing solution. Deeper source could not be localized by these inverse solutions. By the spatial wavelet decomposition prior to source localization, subtopographic maps, which originate from individual sources, can be identified, whose superimposition yields the original ERP topography. In our case, seven octave spatial wavelet decomposition is applied to the topography serie of total activity and the original topography serie is divided into two sets, as seen in Figure 3. One of the topographic map series denotes the approximation outputs, which is the low-pass filtered in all directions, and the other series shows the detail output, which is the remaining part of the original topography series. When these topographic map series are individually localized, it can be observed that the predefined sources can be easily discriminated, as shown in Figure 4. Superficial source is localized from the approximation output while the deeper source is estimated from the detail output. Although it could be decomposed into any number of subtopographies, the original simulated

Figure 3. Topography of superficial source, deeper source and total activity at the time instant of 32; fourth topography is the approximation and the fifth topography is the detail output

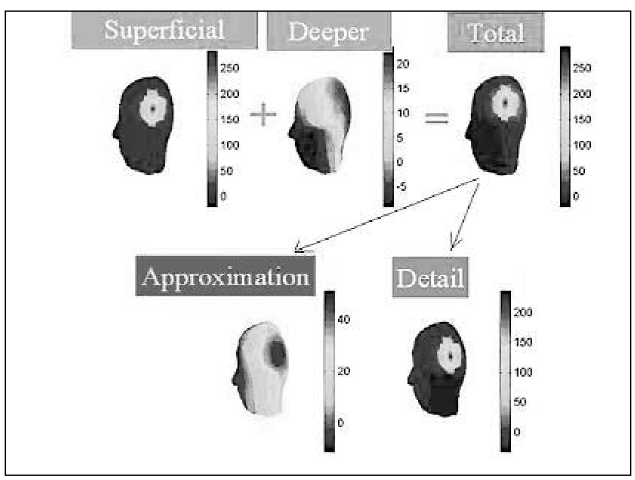
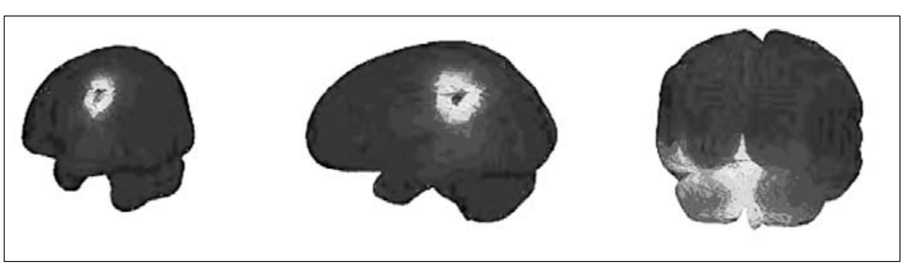


Figure 4. Inverse solutions of the total, approximated and detail topographies with MUSIC, respectively



map is divided into two, each of which corresponds to one individual distinct source.

FUTURE TRENDS

Due to the large number of grid points, the spatial wavelet decomposition requires a high computational complexity and memory usage. More efficient algorithms will significantly decrease the computational cost and memory usage. The mutichannel data projected on the time-frequency and space domains will enhance the localization of nonstationary and distributed sources in the EEG/ERP, and help contribute to overcome the nonuniqueness inverse problem. Implications of these studies are that these individual nonstationary ERP sources when localized in space, time and/or frequency correlate with the distinct and/or interacting cognitive functions of the brain.

CONCLUSION

Spatial wavelet analysis of ERP at a given temporal location yields spatially stationary scalp maps at different spatial frequencies, determined by the depth and extension of individual ERP sources. Therefore, a spatial preprocessing of the ERP simplifies the complexity of the scalp map into several subtopographic maps produced by individual ERP sources prior to their source estimation. This helps us to enhance the accuracy and reliability of the source localization procedures.

REFERENCES

Ademoğlu, A., Demiralp, T., Istefanopulos, Y., Comu, S., & Baykan, B. (2003, September 17–21). Epileptic source localization using wavelet prefiltering and MUSIC scanning. In *Proceedings of the 25th Annual International Conference of the IEEE EMBS* (pp. 2366–2369). Cancun, Mexico.

Baillet, S., Mosher, J. C., & Leahy, R.M. (2001, November). Electromagnetic brain mapping. *IEEE Signal Processing magazine*, 14–30.

Daubechies, I. (1988). Orthonormal bases of compactly supported wavelets. *Comm. in Pure and Applied Math., 41*, 909–996.

Gonzalez Andino, S.L., Grave de Peralta Menendez, R., Lantz, C.M., Blank, O., Michel, C.M., & Landis, T. (2001). Non-stationary distributed source approximation: An alternative to ımprove localization procedures. *Human Brain Mapping, 14*, 81–95.

Grossman, A., & Morlet, J. (1984). Decomposition of hardy functions into square integrable wavelets of constant shape. *SIAM J Math. Anal., 15*, 723–736.

Hamalainen, S.M., & Sarvas, J. (1989). Realistic conductivity geometry model of the human head for interpretation of neuromagnetic data. *IEEE Trans. On Biomedical Eng., 36*(2), 165–171.

Koenig, T., Marti-Lopez, F., & Valdes-Sosa, P.A. (2001). Topographic timefrequency decomposition of the EEG. *NeuroImage, 14*, 383–390.

Koles, Z.J. (1998). Trends in EEG source localization. *Electroenceph. Clin. Neurophysiol., 106*, 127–137.

Mallat, S. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 11*, 674–693.

Meyer, Y. (1990). *Wavelets and operators*. Cambridge University Press.

Michela, C.M., Murraya, M.M., Lantza, G., Gonzaleza, S., Spinellib, L., & Grave de Peraltaa, R. (2004). EEG source imaging. *Clinical Neurophysiology, 115*, 2195–2222.

Miwakeichi, F., Martinez-Montes, E.,Valdes-Sosa, P.A., Nishiyama, N., Mizuhara, H., & Yamaguchia, Y. (2004). Decomposing EEG data into spacetime-frequency components using parallel factor analysis. *NeuroImage, 22*, 1035–1045.

Mosher, J.C., Lewis, P.S., and &, R.M. (1992) Multiple dipole modeling and localization from spatio-temporal MEG data. IEEE Trans. *On Biomedical Engineering, 39*(6), 541–557.

Pascual-Marqui, R.D., Michel, C.M., & Lehmann, D. (1994). Low resolution electromagnetic tomography: A new method for localizing electrical activity in the brain. *Int. J. Psychophysiol., 18*, 49–65.

Scherg, M., & von Cramon, D. (1986). Evoked dipole source potentials of the human auditory cortex. *Electroencephalography and Clinical Neurophysiology, 65*, 344–360.

Schlitt, H., Heller, A.L., Aaron, R., Best, E., & Ranken, D.M. (1995). Evaluation of boundary element methods for EEG forward problem: Effect of linear interpolation. *IEEE Trans. On. Biomedical Eng., 42*(1), 52–57.

Wang, K., Begleiter, H., & Porjesz, B. (1998). Spatial enhancement of eventrelated potentials using multiresolution analysis. *Brain Topogr. Spring, 10*(3), 191–200.

Zhukov, L., Weinstein, D., & Johnson, C. (2000, May/June). Independent component analysis for EEG source localization an algorithm that reduces the complexity of localizing multiple neural sources. *IEEE Engineering in Medicine and Biology*, 87–96.

KEY TERMS

Boundary Element Method (BEM): BEM is a numerical computational method of solving linear partial differential equations, which have been formulated as integral equations.

Event-Related Potential (ERP): ERP is any measured brain response that is directly the result of a thought or perception.

Forward Problem of EEG: The procedure that calculates the electrical potential over the scalp.

Inverse Problem of EEG: The procedure which estimates the activity of the brain from given electrical potential data.

Multiresolution Decomposition (MRA): MRA is the design method of most of the practically relevant discrete wavelet transforms (DWT), and the justification for the algorithm of the fast wavelet transform (FWT).

Spatial Frequency: The spatial frequency is a measure of how often the structure repeats per unit of distance.

Subtopography: Simple scalp topography originated from decomposition of main scalp map.

Spatial Heart Simulation and Adaptive Wave Propagation

S

Sándor Miklós Szilágyi
Hungarian Science University of Transylvania, Romania

László Szilágyi
Hungarian Science University of Transylvania, Romania

Attila Frigy
County Medical Clinic Nr. 4, Târgu Mures, Romania

Levente Károly Görög
Hungarian Science University of Transylvania, Romania

Zoltán Benyó
Budapest University of Technology and Economics, Hungary

INTRODUCTION

Sudden cardiac death, caused mostly by ventricular fibrillation, is responsible for at least five million deaths in the world each year. Despite decades of research, the mechanisms responsible for ventricular fibrillation are not well understood. As most computational studies are limited primarily to planar simulations, experiments so far have not elucidated the mechanisms responsible for spatial phenomenon (Janse, Wilms-Schopman, & Coronel, 1995) of ventricular fibrillation.

It would be important to understand how the onset of arrhythmias that cause fibrillation depends on details such as heart size (Winfree, 1994), geometry (Vetter & McCulloch, 1998; Panfilov, 1999), mechanical and electrical state, anisotropic fiber structure (Fenton & Karma, 1998), and inhomogeneities (Antzelevitch et al., 1999; Wolk, Cobbe, Hicks, & Kane, 1999). The main difficulty in development of a quantitatively accurate simulation of an entire three-dimensional human heart is that human heart muscle is a strongly excitable medium whose electrical dynamics involve rapidly varying, highly localized fronts (Cherry, Greenside, & Henriquez, 2000).

In ventricular tissue (which is the most important to study) the width of a depolarization front is usually less than half mm and a simulation that approximates the dynamics of such a front requires a spatial resolution of $\Delta x \leq 0.1$ mm. Forasmuch the muscle in an adult heart has a volume of 250 cm^3, and so a uniform spatial representation requires at least $2.5 \cdot 10^8$ nodes. Taking into account that each node's state is described with at least 50 floating number (with at least 4 byte resolution) the necessary storage space rises higher than 50GB (personal computers have no opportunity to handle such a huge amount of data in memory). It is known that the rapid depolarization of the cell membrane (this is the fastest event in heart functioning) blow over in few hundred microseconds. A reasonable resolution of this event requires a time step $\Delta t \leq 25$ microseconds. Since dangerous arrhythmias such as ventricular fibrillation may require several seconds to become established, the 10^{10} floating point numbers associated with the spatial representation would have to be evolved over 10^5 - 10^6 time steps. Such a huge uniform mesh calculation currently exceeds all existing computational resources (Cherry, Greenside, & Henriquez, 2003).

Previous efforts to improve the efficiency of simulations have followed several approaches. The most common way of simplification is based on a reduced mathematical model that can reproduce some of the behavior observed in more complex models but with only a few coupled fields. One widely used example is the two-variable FitzHugh–Nagumo model (FitzHugh, 1961), which describes behavior of a general excitable medium. This model can be modified to approximate various types of cardiac dynamics (Berenfeld & Pertsov, 1999). Another example is a three-variable model developed by Fenton and Karma (Fenton & Karma, 1998) that is designed to reproduce the restitution curves of more complex cardiac models. Albeit these and other simplified models can reproduce many known features of cardiac electrical propagation, some electronic (Fenton, 2000) and drug effects cannot be handled properly. Accordingly, efficient algorithms for more quantitatively-based models are still desirable.

The spatiotemporal structure of wave dynamics in excitable media suggests an automatically adjustable resolution in time and space. The basic idea of this improvement (Cherry et al., 2000, 2003) is deducted from experiments (Winfree et al., 1996; Witkowski et al., 1998) and simulation (Courtemanche, 1996), which recommends that the function of electrical membrane potential of a ventricular cell $f_V(t,x,y,z)$ in the fibrillating state consists of many spirals (for approximately two-dimensional tissue such as the right ventricle and atrium) or of many scroll waves (for thicker cardiac tissue such as the left ventricle). An interesting property of these spatiotemporal disordered states is that the dynamics is *sparse*: at any given moment, only a small volume fraction of the excitable medium is depolarized by the fronts, and away from them the dynamics are slowly varying in space and time. This idea permits the decrement of necessary computational effort and storage space (only the time dependent depolarization fronts have to be represented with high resolution instead of the whole tissue volume) for regular beats but the total front volume can greatly increase with a fibrillating state. By varying the spatiotemporal resolution to concentrate computational effort primarily along the areas with large spatial and temporal gradients, it is possible to reduce greatly the computational effort and memory required.

Most of previous studies (Rousseau & Kapral, 2000; Vigmond & Leon, 1999) implemented the adaptivity, locally and dynamically, by varying either the spatial or temporal resolution, but not both. However, these methods decreased the global processing volume significantly and they generally could not yield computational savings as significant as the time- and space-adaptive strategies. All of these adaptive approaches have the advantages that they are largely model- and method-independent.

The heart modeling method presented in this article is based on an algorithm developed by Berger and co-workers (Bell, Berger, Saltzman, & Welcome, 1998) and used successfully for three-dimensional simulations. Although the earlier presented methods involved dynamic time and space analysis, the a priori heart irregularities were neglected. The Purkinje fiber net has an unknown shape and highly varies from patient to patient. To handle this problem, we involved in our method the average heart tissue-topology that admits the appliance of probabilistic descriptions.

The rest of the article describes the applied human cell and tissue model, the time and spatial position dependent heart and torso model, the adaptively variable resolution wave-propagation method, and the parallel processing of these algorithms aided by graphic cards.

MATERIALS AND METHODS

Applied Human Cell and Tissue Model

We used the ten Tusscher heart cell model (ten Tusscher, Noble, Noble, & Panfilov, 2004) for ventricular and Nygren's model (Nygren, Fiset, Firek, Clark et al., 1998) for atrial cells, to investigate the accuracy and efficiency of the simulation algorithm. These models are based on recent experimental data on most of the major ionic currents, such as the fast sodium, L-type calcium, transient outward, rapid and slow delayed rectifier, and inward rectifier currents. With the inclusion of basic calcium dynamics, the contraction and restitution mechanism of the muscle cells can be investigated.

The model is able to reproduce human epicardial, endocardial, and M cell action potentials, to modify the internal state of the cells and to show that differences can be explained by differences in the transient outward and slow delayed rectifier currents. These properties allow the study of evolvement of reentrant arrhythmias. The conduction velocity restitution of this model is broader than in other models and agrees better with available data. We conclude that the applied model can reproduce a variety of electrophysiological behaviors and provides a basis for studies of reentrant arrhythmias in human heart tissue.

As presented in ten Tusscher et al. (2004), the cell membrane is modeled as a capacitor connected in parallel with variable resistances and batteries representing the different ionic currents and pumps. The electrophysiological behavior of a single cell can be described by equation (1):

$$\frac{dV}{dt} = -\frac{I_{ion} + I_{stim}}{C_{memb}} \tag{1}$$

where V is the voltage, t is time, I_{ion} is the sum of all trans-membrane ionic currents, I_{stim} is the externally applied stimulus current, and C_{memb} is the cell capacitance per unit surface area.

The ionic current's sum is given by the equation (2):

$$I_{ion} = I_{Na} + I_{K1} + I_{to} + I_{Kr} + I_{Ks} + I_{CaL} + \\ + I_{NaCa} + I_{NaK} + I_{pCa} + I_{pK} + I_{bCa} + I_{bK} \quad (2)$$

where I_{NaCa} is Na^+/Ca^{2+} exchanger current, I_{NaK} is Na^+/K^+ pump current, I_{pCa} and I_{pK} are plateau-, I_{bCa} and I_{bK} are background- Ca^{2+} and K^+ currents. The fast Na^+ current that is responsible for the fast depolarization of the cardiac cells is formulated by the equation (3):

$$I_{Na} = G_{Na} \cdot m^3 \cdot h \cdot j \cdot (V - E_{Na}) \quad (3)$$

where G_{Na} is the sodium conductance, *m* represents the activation gate, *h* the fast- and *j* the slow- inactivation gate. All detailed equations are described in Harrild and Henriquez (2000). These gates have mainly a voltage dependent behavior. The maximal value of the first derivative of the L-type calcium current I_{CaL}, transient outward current I_{to}, slow delayed rectifier current I_{Ks}, rapid delayed rectifier current I_{Kr} and inward rectifier K^+ current I_{K1} and all other described currents is lower with at least two magnitudes than for the fast Na^+ current I_{Na}.

A homogenous spatial cardiac tissue can be modeled in space as a continuous system with the following partial differential equation (4) (See Box 1) where ρ_x, ρ_y, ρ_z are the cellular resistivity and S_x, S_y, S_z are the surface-to-volume ratio in the *x*, *y* and *z* directions.

Computational modeling of the cardiac tissue is a useful tool for developing mechanistic insights into cardiac dynamics. The most important parts of human cardiac analysis are atria and ventricular tissue modeling.

Our study is based on Harrild's atria model (Harrild & Henriquez, 2000) that is the first membrane-based description of spatial conduction in a realistic human atrial geometry. This model includes both the left and right atria, including representations of the major atrial bundles and a right-sided endocardial network of pectinate muscle. The membrane's kinetics is governed by the Nygren et al. (1998) formulation for the human atrial cell. An advantage of this model is that it provides an easy perceptibility of atrial activation, particularly in regions that cannot be easily recorded in patients.

It has long been appreciated that cardiac ventricular fibers are arranged as counter-wound helices encircling the ventricular cavities, and that orientation of these fibers depends on transmural location (Streeter & Hanna, 1973). Fibers tend to lie in planes parallel to the epicardium, approaching a longitudinal orientation on the ventricular surfaces, and rotating toward the horizontal near the midwall. The direct anatomical reconstructions are labor-intensive and time-consuming tasks. In our study we applied Winslow's ventricular tissue model (Winslow, Scollan, Holmes et al., 2000).

Heart and Torso Model

There are many possible different heart structures (Quan & Evans, 1998). To describe various representative cases, we studied our breast MRI records (42 examples) and numerous CT images. These samples lead us to construct a morphological heart structure for simulation. The anatomical structure of the atria (Harrild & Henriquez, 2000) and ventricles (ten Tusscher, Bernus, Hren, & Panfilov, 2006; Vetter & McCulloch, 2001) was involved in geometrical model of the heart and torso, which was constructed using tetra meshes. The torso, lung, endo- and epicardial surfaces were divided into 23647, 38844, 78809, and 89723 subunits. Such a heart model could have a maximal spatial resolution of 0.025 mm that means more than ten billion individual compartments at highest decomposition. For a flexible simulation, we could choose the minimal time-slice between 0.01ms and 2ms over the whole ventricular excitation cycle.

The structure of the torso, its position in space, the relative position and distance of the compartments with respect to the electrodes, and the electrical behaviour of the torso's contents are necessary to be known (Ramon, Czapski, Haueisen et al., 1998). As the model has to take in consideration extremely numerous parameter values, the problem cannot be solved in a determin-

Box 1.

$$\frac{dV}{dt} = \frac{1}{C_{memb}} \cdot \left(-I_{ion} - I_{stim} + \frac{1}{\rho_x S_x} \cdot \frac{\partial^2 V}{\partial x^2} + \frac{1}{\rho_y S_y} \cdot \frac{\partial^2 V}{\partial y^2} + \frac{1}{\rho_z S_z} \cdot \frac{\partial^2 V}{\partial z^2} \right) \quad (4)$$

istic way (we have much more unknown values than known equations). That is why a stochastic method (genetic algorithm, adaptive neural networks, and fuzzy systems) should be applied to determine the values of the parameters.

Adaptively Varied Resolution and Parallel Implementation

The adaptive algorithm assumes that some user specified explicit relation has been chosen to approximate both time and space derivatives. For each subunit, the spatial and temporal resolution was determined using the calculated derivatives. The established coarse or fine level assures that internal simulated data can be interpolated with the desired precision from input information. As temporal units vary with the spatial resolution, more time steps are needed for finer patches, which lead to an asynchronous updating. Data on all resolution levels are synchronized only after one full time step on the coarsest grid level is completed. The efficiency of the method arises from its ability to refine or to coarsen the spatial and temporal representations of subunits automatically and locally. The approximation errors are estimated on each subunits to determinate the lowest necessary resolution to keep under a predefined tolerance value. This method is highly parallelizable. As subunit values are determined independently, these tasks can be effectuated on separate processors with reduced communication needs.

The hardware accelerated programmability of graphics processing units (GPUs) admits the development of programs called shaders. These programs, called vertex and fragment shaders, are loaded into the graphics card memory for replacing the fixed functionality. The fragment shaders can be used in our method to perform the single-instruction-multiple-data (SIMD) commands for each subunit. From architectural concepts result that the GPUs are most efficient in case of more than 1000 similar tasks.

Simulation Platforms

Experiments have been performed on four platforms with different configuration that are described in Table 1. Each computer has 1GB memory that admits a maximal resolution between 0.01 mm (normal beats) and 0.1 mm (ventricular fibrillation) for the most critic area (highly restricted size).

RESULTS

Table 2 informs us about simulation times using various configurations. All simulation tasks were performed with adaptive time and spatial resolution, and Table 2 show the finest ones among them. The simulated normal and pathological cases have one second duration. In all cases, the number of simultaneously performable tasks has an order of thousands. The conventional simulation (constant resolution) was performed only for 1 mm spatial and 0.2 ms temporal units, and was slower about 200 times in normal and 35 times in fibrillating case.

Section a of Figure 1 elucidates the relation between estimation error and spatial resolution. The obtained results are almost similar for normal and pathological cases.

In section b of Figure 1, we can observe the propagation of depolarization wave in an anisotropic ventricular tissue. The studied surface contain a rotated tissue, where the ratio of the conduction speed for an almost horizontal and vertical direction is considered 3.2.

Figure 2 presents a cell activation simulation that was effectuated and visualized after 200ms of the rises of sinoatrial node excitation. In the right part of

Table 1. Configuration of the simulation platforms involved in this study

Configuration	CPU	GPU	Memory
1st	Athlon 3000+	Nvidia 6600	1GB DDR
2nd	Core2 Duo 6400	Ati 1950 PRO	1GB DDR2
3rd	Core2 Duo 6400	2x Ati 1950 PRO	1GB DDR2
4th	Pentium D 805	Nvidia 7600GT	1GB DDR2

the image, the activation time of a cell is visualized. Each slice has 10mm distance from other, so totally a five centimeter wide ventricular tissue is visualized. Figure 3 presents a normal and pathological simulated 12 lead ECG signal.

DISCUSSION AND CONCLUSION

Table 1 present four configurations with shader model (SM) 3.0 ready GPUs. The third configuration is the most powerful due to the cross-fire connected ATI 1950 PROs. The type of CPU (Intel or AMD), the clock speed (1.86 - 2.66 GHz), the core number (solo or duo), and memory bandwidth (DDR or DDR2) did not play an important role because a powerful video card has a much higher floating-point calculation power (internally has 8-36 pixel shader units). In all cases, the size of memory was selected at 1GB that restricts the applicable maximal resolution.

From data of Table 2, we can observe the clear dominance of GPUs. Although the spatial and temporal resolution limit the necessary simulation time in all cases a massive parallelization could be performed.

Table 2. The simulation time of a one second duration event performed on involved platforms using two spatial and temporal resolutions. In simulation a normal beat, ectopic beat, and ventricular fibrillation state was included.

Conf. & resolution/ Pathological case	**Normal beat**	**Ectopic beat**	**Ventricular fibrillation**
1st – (1 mm, 0.2 ms)	11.3 sec	37.15 sec	4 min 2 sec
2nd – (1 mm, 0.2 ms)	1.32 sec	5.21 sec	33.11 sec
3rd – (1 mm, 0.2 ms)	0.7 sec	2.68 sec	17.13 sec
4th – (1 mm, 0.2 ms)	2.37 sec	9.03 sec	57.48 sec
1st – (0.1 mm, 0.05 ms)	1h 11 min	4h 22 min	1d 5h 11 min
2nd – (0.1 mm, 0.05 ms)	9 min 20 sec	37 min 10 sec	3h 53 min
3rd – (0.1 mm, 0.05 ms)	5 min 3 sec	19 min 17 sec	1h 59 min
4th – (0.1 mm, 0.05 ms)	15 min 40 sec	1h 6 min	6h 42 min

Figure 1. (a) The relation between the applied spatial resolution and resulting estimation error simulating a normal beat. (b) The simulated depolarization wave in anisotropic ventricular tissue is presented. From the pace origin the wave front propagation is plotted with 5ms resolution.

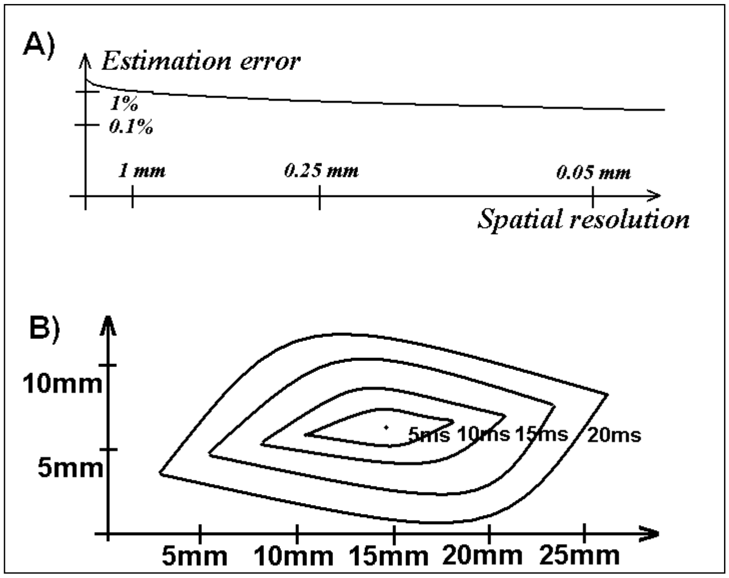
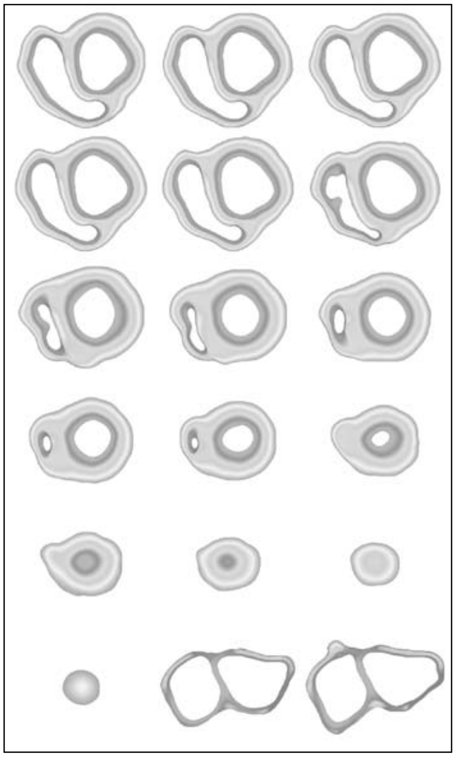


Figure 2. The simulated depolarization wave propagation in ventricles and atria in case of normal QRS beat



All shader programs were created using a low level programming environment. We could observe that in normal cases the active depolarization wave front has a much lower size than in case of ventricular beat of ventricular fibrillation. In a complex biological situation, as the wave front size grows the parallelization becomes harder. This assumption is reflected by the simulation times from Table 2. It is observable that a normal heart has at least 20 times lower front area than a fibrillating one. As a cardiac muscle (especially left ventricular) becomes less homogeny, the relative simulating speed decreases. Some basic characteristics of the heart such as size, maximal tissue volume, and left wall width significantly influence the maximal performance.

Section a of Figure 1 represent the estimation error in function of spatial resolution. The temporal resolution has almost similar effect but with lower impact. From measurements, we could deduct that estimation error is physiological state free. In normal and pathological cases, we measured almost the same error values.

In this article, we have discussed new features and new capabilities of a space-time adaptive heart-modeling algorithm. We have shown the algorithm's ability to simulate inhomogeneous and strongly anisotropic tissue regions.

This method can provide a variety of advances in addition to reductions in time and memory requirements. For example, the algorithm can allow a more complex ionic model and higher spatial resolution of a nonlinear tissue model. Similarly, it can allow the use of higher spatial and temporal resolution to reduce the angle dependence of propagation patterns in spatial domains with rotational anisotropy or to verify that a calculation is sufficiently resolved so that an increase in resolution does not affect the results (see section a of Figure 1).

We have presented a massively parallelized flexible and efficient heart simulation method that uses almost all features of a modern processing hardware. After that, we demonstrated that the processor of a modern graphics card can afford better performance than a modern CPU under certain conditions, in particular, allocating data in a regular and parallel manner. In these situations, the GPU should operate in a SIMD fashion to get the most performance hit. Experimental results show that the graphics card can be exploited in order to perform nonrendering tasks.

ACKNOWLEDGMENT

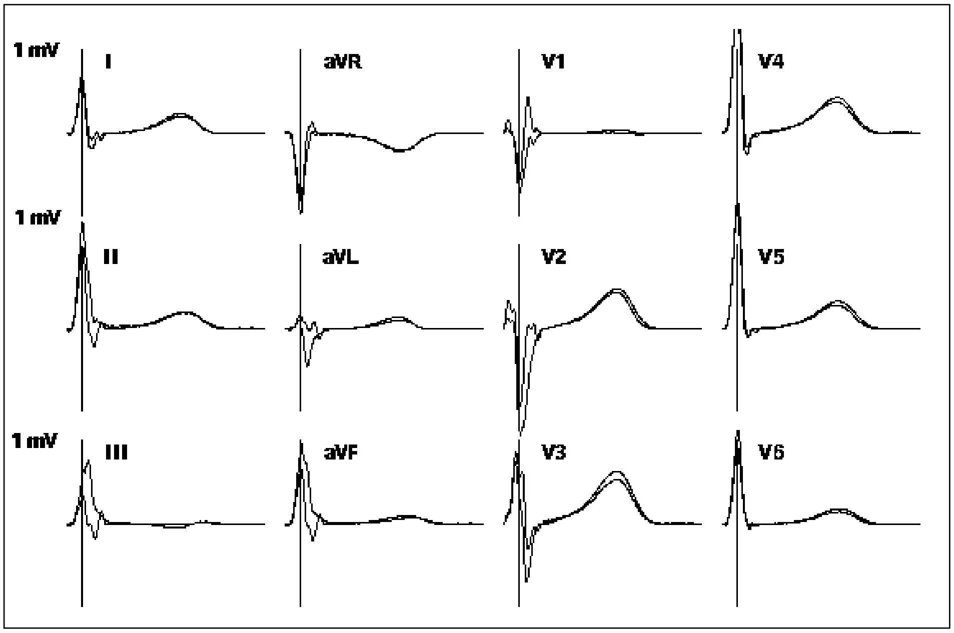
This research was supported in part by the Sapientia Institute for Research Programmes, Domus Hungarica Scientiarium et Artium, and the Communitas Foundation.

REFERENCES

Antzelevitch, C., Shimizu, W., Yan, G.-X., Sicouri, S., Weissenburger, J., Nesterenko, V.V., et al. (1999). The M cell: Its contribution to the ECG and to normal and abnormal electrical function of the heart. *J. Cardiovasc. Electrophysiol., 10,* 1124-1152.

Bell, J., Berger, M., Saltzman, J., & Welcome, M. (1994). Three dimensional adaptive mesh refinement for hyperbolic conservation laws. *SIAM J. Sci. Comput., 15*, 127-138.

Figure 3. The simulated ECG signal in normal and abnormal case



Berenfeld, O., & Pertsov, A.M. (1999). Dynamics of intramural scroll waves in three-dimensional continuous myocardium with rotational anisotropy. *J. Theor. Biol., 199,* 383-394.

Cherry, E.M., Greenside, H.S., & Henriquez, C.S. (2000). A space-time adaptive method for simulating complex cardiac dynamics. *Phys. Rev. Lett., 84,* 1343-1346.

Cherry, E.M., Greenside, H.S., & Henriquez, C.S. (2003). Efficient simulation of three-dimensional anisotropic cardiac tissue using an adaptive mesh refinement method. *Chaos, 13,* 853-865.

Courtemanche, M. (1996). Complex spiral wave dynamics in a spatially distributed ionic model of cardiac electrical activity. *Chaos, 6,* 579-600.

Fenton, F. (2000). Numerical simulations of cardiac dynamics. What can we learn from simple and complex models? *Comput. Cardiol., 27,* 251-254.

Fenton, F., & Karma, A. (1998). Vortex dynamics in 3D continuous myocardium with fiber rotation: Filament instability and fibrillation. *Chaos, 8,* 20-47.

FitzHugh, R.A. (1961). Impulses and physiological states in theoretical models of nerve membranes. *Biophys. J., 1,* 445-466.

Harrild, D.M., & Henriquez, C.S. (2000). A computer model of normal conduction in the human atria. *Circulation Research, 87,* 25-36.

Janse, M.J., Wilms-Schopman, F., & Coronel, R. (1995, July). Ventricular-fibrillation is not always due to multiple wavelet reentry. *J. Cardiovasc. Electrophysiol., 6,* 512-521.

Nygren, A., Fiset, C., Firek, L., Clark, J.W., et al. (1998). Mathematical model of an adult human atria cell: The role of K+ currents in repolarization. *Circulation Research, 82,* 63-81.

Panfilov, A.V. (1999). Three-dimensional organization of electrical turbulence in the heart. *Phys. Rev., E(59),* R6251-R6254.

Quan, W., & Evans, S.J. (1998). Efficient integration of a realistic two-dimensional cardiac tissue model by domain decomposition. *IEEE Trans. Biomed. Eng., 45,* 372-384.

Ramon, C., Czapski, P., Haueisen J., et al. (1998). MCG simulations with a realistic heart-torso-model. *IEEE Trans. Biomed. Eng., 45,*1323-1331.

Rousseau, G., & Kapral, R. (2000). Asynchronous algorithm for integration of reaction-diffusion equations for inhomogeneous excitable media. *Chaos, 10*, 812-825.

Streeter, D.D., & Hanna, W.T. (1973). Engineering mechanics for successive states in canine left ventricular myocardium: II. Fiber angle and sarcomere length. *Circulation Research, 33*, 656-664.

ten Tusscher, K.H.W.J., Bernus, O., Hren, R., & Panfilov, A.V. (2006, January-April). Comparison of electrophysiological models for human ventricular cells and tissues. *Prog Biophys Mol Biol., 90(1-3)*, 326-345.

ten Tusscher K. H. W. J., Noble D., Noble P. J., Panfilov A. V. (2004) A model for ventricular tssue. *Am J Physiol Heart Circ Physiol 286*, H1573-H1589.

Vetter, F.J., & McCulloch, A.D. (1998). Three-dimensional analysis of regional cardiac function: A model of the rabbit ventricular anatomy. *Prog. Biophys. Mol. Biol., 69*, 157-183.

Vetter, F.J., & McCulloch, A.D. (2001, May). Mechanoelectric feedback in a model of the passively inflated left ventricle. *Ann Biomed Eng., 29(5)*, 414-426.

Vigmond, E.J., & Leon, L.J. (1999). A computationally efficient model for simulating electrical activity in cardiac tissue with fibre rotation. *Ann. Biomed. Eng., 27*, 160-170.

Winfree, A.T. (1994). Electrical turbulence in three-dimensional heart muscle. *Science, 266,* 1003-1006.

Winfree, A.T., et al. (1996). Quantitative optical tomography of chemical waves and their organizing centers. *Chaos, 6*, 617-626.

Winslow, R.L., Scollan, D.F., Holmes, A., et al. (2000). Electrophysiological modeling of cardiac ventricular function: From cell to organ, Annual Review. *Biomed Eng., 2*, 119-155.

Witkowski, F.X., et al. (1998). A method for visualization of ventricular fibrillation: Design of a cooled fiberoptically coupled image intensified CCD data acquisition system incorporating wavelet shrinkage based adaptive filtering. *Chaos, 8,* 94-102.

Wolk, R., Cobbe, S.M., Hicks, M.N., & Kane K.A. (1999). Functional, structural, and dynamic basis of electrical heterogeneity in healthy and diseased cardiac muscle: implications for arrhythmogenesis and anti-arrhythmic drug therapy. *Pharmacol. Ther., 84,* 207-231.

KEY TERMS

Cell Model: A model that describes the electric and mechanic functioning of a cardiac cell.

Graphics Processing Unit (GPU): A dedicated graphics rendering device for a personal computer, workstation, or game console.

Heart Model: A model that describes the electric and mechanic functioning of the whole heart.

Heart Simulation: A simulation that focuses on the functional modeling of the heart.

Inverse Electrocardiography: A methodology that reconstructs the inner structure of the heart using multichannel ECG.

Space–Time Adaptive Mesh Refinement Algorithm: An adaptive mesh refinement method, able to vary the spatial and temporal resolution.

Torso Model: A model that describes the electric connections among heart regions and surface electrodes.

S

Spatial Heart Simulation and Analysis Using Unified Neural Network

Sándor Miklós Szilágyi
Hungarian Science University of Transylvania, Romania

László Szilágyi
Hungarian Science University of Transylvania, Romania

Zoltán Benyó
Budapest University of Technology and Economics, Hungary

INTRODUCTION

The most important health problem affecting large groups of people is related to the malfunction of the heart, usually caused by heart attack, rhythm disturbances, and pathological degenerations. One of the main goals of health study is to predict these kinds of tragic events, and by identifying the patients situated in the most dangerous states, to make it possible to apply a preventing therapy.

Creating a heart model is important (Thaker & Ferrero, 1998) as the computer, while applying traditional signal processing algorithms recognizes lots of waves, but it does not really "understand" what is happening. To overcome this, the computer needs to know the origin and the evolvement process of the ECG signal (MacLeod & Brooks, 1998). During signal processing, if the traditional algorithm finds an unrecognizable waveform, the model-based approach is activated, which tries to estimate the causes of the encountered phenomenon (e.g., quick recognition of ventricular fibrillation) (Szilágyi, 1998).

The heart is a dynamic organ and places special demands on spatial techniques. To understand its physiology and pathophysiology, not only the electrical activity and spatial distribution of its structures is important, but also their movement during cardiac cycles. The measured ECG signal is influenced during repolarization by the mechanical movement of the heart.

The main goal of the inverse problem of ECG is to characterize and reconstruct cardiac electrical events from measurements. In contrast to the forward problem of electrocardiography, the inverse problem does not possess a mathematically unique solution and in order to improve stability, it needs to adopt regularization techniques.

Several approaches have been explored to handle the problem of multiple solutions by using equivalent cardiac generators (such as equivalent dipole and multipole), heart surface isochrones (Cuppen & Oosterom, 1984), or epicardial potential (Guanglin & Bin, 2001). The high sensitivity of solutions to the different disturbances forced the investigators to explore regularization techniques (Shahidi, Savard, & Nadeau, 1994). These methods allow a significant progress, but the different uncertainty elements of the processing hinder the potentially beneficial ECG inverse solutions from becoming a routine clinical tool at present.

Body surface potential mapping (BSPM) was developed to allow an almost complete data acquisition from the body surface (Mirvis, 1988). BSPM may have a great advantage over the standard 12-lead system in different situations due to deeper accessible information. Mirvis (1988) has shown some cases of BSPM recordings that clearly demonstrate the inadequacies of the standard ECG lead sets in a variety of pathologies. As we know more about the depolarization-repolarization mechanism, we can understand better the internal function of the heart.

Many interesting biomedical applications of artificial neural networks are in the area of data processing (Minami, Nakajima, & Yoyoshima, 1999). The best known neural solutions involve multilayer perceptrons, Kohonen self-organizing networks, fuzzy or neuro-fuzzy systems, genetic algorithms, and the combination of various solutions within a hybrid system (Lagerholm, Peterson, Braccini, Edenbrandt, & Sornmo, 2000). A typical heart modeling system applies many neural networks and chooses the best one, while discarding the rest. Most efficient approaches usually are based on the combination of many classifiers utilizing either different classifier network structures or different data preprocessing methods (Osowski & Hoai, 2001).

The support vector machine (SVM), pioneered by Vapnik (1998) assumed to solve the main drawbacks of conventional artificial neural networks (ANNs) such as:

- Modern biological problems are high-dimensional, and if the underlying mapping is not very smooth, the linear paradigm needs an exponentially increasing number of terms with an increasing dimensionality of the input space, which means an increase in the number of independent variables. This is known as "the curse of dimensionality."
- The real-life data generation laws may typically be far from the normal distribution and a model-builder must handle this difference in order to construct an effective learning algorithm.
- The maximum likelihood estimator (and consequently the sum-of-error-squares cost function) should be replaced by a new induction paradigm that is uniformly better, in order to model properly non-Gaussian distributions.

The SVM classifiers became quite popular due to their robustness and stability (Osowski, Hoai, & Markiewicz, 2004). An SVM used in a heart modeling system is rigorously based on statistical learning theory and simultaneously minimizes the training and test errors. Apart from that, they produce a unique globally optimal solution and hence are extensively used in diverse applications including medical diagnosis (Smola & Cholkopf, 1998).

This article presents an event recognition study performed with ECG signal analysis and 3D heart model using unified neural networks (UNN). The main purpose is to evaluate the strength and weakness of each method and to analyze the cooperation efficiency in malfunction diagnosis.

MATERIALS AND METHODS

Unified Neural Networks

If we focus on the two-class classification case and consider linear discriminant functions, the respective decision hypersurface in the *n*-dimensional feature space is a hyperplane that can be described as:

$$g(x) = w^T \cdot x + w_0 = 0 \qquad (1)$$

where $w = [w_1, ..., w_n]^T$ is known as the weight vector and w_0 as the threshold value. For a given vector x_d, if the function $g(x_d) = 0$ than x_d situates on the decision hyperplane. The distance z of a vector x from the decision hyperplane can be represented as: $z = \frac{|g(x)|}{|w|}$, where $|w| = \sqrt{\sum_{i=1}^{n} w_i^2}$.

In a classification problem, our purpose is the optimization of vector w in such a way, that the criteria function $J(w)$ is minimized. Let ω_1 and ω_2 be the two classes that we have to separate. We assume that this task can be performed using a linear relation. This means that there exists at least one such a w_{sol} hyperplane that fulfill the following relations:

$$w_{sol}{}^T \cdot x > 0 \text{ for } \forall x \in \omega_1$$
$$w_{sol}{}^T \cdot x < 0 \text{ for } \forall x \in \omega_2. \qquad (2)$$

If we design a classifier, where the desired output is $y_{des} = +1$ for $\forall x \in \omega_1$ and $y_{des} = -1$ for $\forall x \in \omega_2$, and try to modify weights in vector w in such a way that the criteria function $J(w) = (y_{des} - x^T \cdot w)^2$ is minimized, than we have constructed a root mean square error based separator method.

A very popular classifier algorithm is based on SVM-s. The main concept incorporates the search for the "most robust solution" vector w_{sol} that gives the maximum possible margin. The margin is represented by the minimal distance:

$$z = \frac{|g(x)|}{\|w\|}$$

This means the minimization of $\|w\|$.

Although these methods deliver good results in a certain noise-free environment, in biomedical simulation such sterile conditions never occur. The main reason of this problem is caused by measurement errors and the improper estimation of unmeasurable biological parameters.

The root mean square classifiers have the following drawbacks:

- Improper solution in case of asymmetric transfer functions
- Large estimation error of the criteria function in case of border-close high dispersion (uncertain) inputs
- In a noisy environment, the criteria function may possess multiple local minimal solutions that may cause low quality results
- The white noise is "unlearnable" so the function $J(w)$ will saturate at an uncontrollable error level.

The SVM produces a much better result in a hostile environment, but did not take into consideration the topology of the input vectors.

To overcome the above mentioned problems, we propose for classification a UNN. The main difference between unified neural network (UNN) and the described classifiers consists in the equation of criteria function:

$$J(w) = \lambda_d \cdot (y_{des} - x^T \cdot w)^2 + \lambda_{sm} \cdot \sum_{i=1}^{N} f_{sm}(x^T \cdot w) + \lambda_m \cdot f_m(z) \quad (3)$$

that is composed by a difference error, smoothness and margin part. Coefficients λ_d, λ_{sm}, λ_m are weighting the square error and the results of smoothing and margining functions represented by f_{sm} and f_m. The margin value is represented by $z = \frac{|g(x)|}{\|w\|}$.

Although this example illustrates a linear separation, the UNN may be applied in nonlinear environment.

Study Records

The first signal resource was a 192-electrode measurement (BSPM) database (sampled at 1000 Hz with 12-bit resolution) obtained from the Research Institute for Technical Physics and Materials Science (MTA-MFA) of Budapest. These registrations contain various malfunction cases as WPW syndrome, atrial and ventricular fibrillation, and flutter. In the second stage of the study, we used 12-lead ECG registrations from our database. These signals were sampled at 500-1000 Hz with 12-bit resolution.

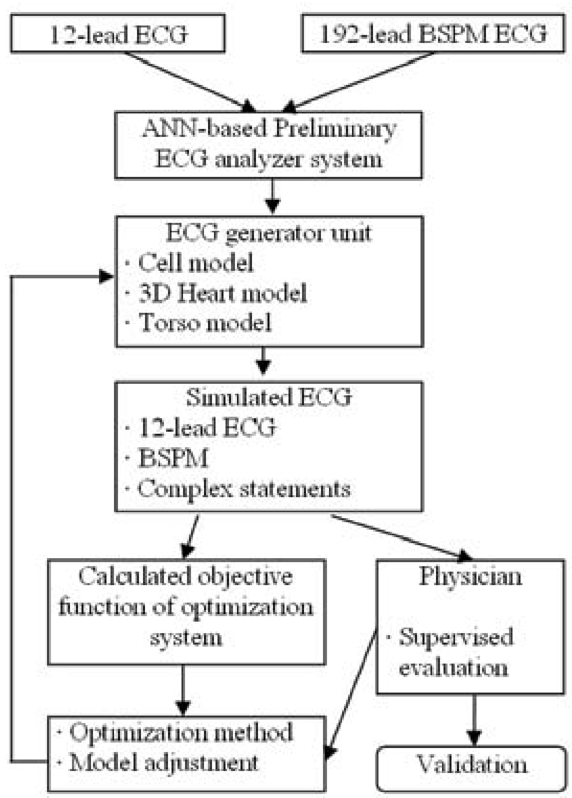
The Approach of ECG Inverse Problem

In contrast to methods that directly solve the matrix equation linking electrical sources with electrical potential fields to estimate ECG inverse solution, our approach indirectly obtains the result in terms of heart model parameters.

The preliminary ECG analyzer system (PAS) is based on detailed, a priori knowledge of human anatomy and physiology, developed using an ANN, tested and validated by physicians in clinical environment (see Figure 1).

In this study, the PAS was used to obtain initial information on the site of origin of cardiac activation. The output of the ANN provides the initial heart model parameters. Then the BSPMs or 12-lead ECGs

Figure 1. The schematic diagram of the heart analyzer and modeling method



were simulated by the ECG generator unit, and the objective functions that assess the similarity between the measured and simulated signals were determined. The heart model parameters are adjusted with the aid of optimization algorithms or in certain cases physicians. The simulation procedure is performed until the objective functions satisfy the given convergence criteria. Finally the parameters are validated.

ANN-Based Preliminary ECG Analyzer System

Application of a priori knowledge to reduce the searching space of heart model parameters is quite important. The PAS (see Figure 2) was developed to determine roughly the cardiac status and state, which was then used to initialize the model parameters and decrease the searching space for the optimization system.

In the present study, the PAS was developed based on a three-layer UNN as shown in Figure 2. This network is capable of mapping the nonlinear input-output relationship to a desired degree of accuracy. An adaptively weighted coefficient algorithm was used to train the ANN. The number of neurons in the input layer was set to 192, corresponding to the number of body surface electrodes used in the present simulation study. The hidden layer had 125 neurons that were determined heuristically. The output layer had 512 neurons, which corresponded to 32 ventricular myocardial segments of computer heart model: *MS-1* to *MS-32*. Sixteen cardiac cellular units were selected for each of 32 myocardial segments in the ventricles, and each of these 16 x 32 = 512 sites was then paced in the forward simulation using the computer heart-torso model, generating the dataset for training the ANN.

RESULTS

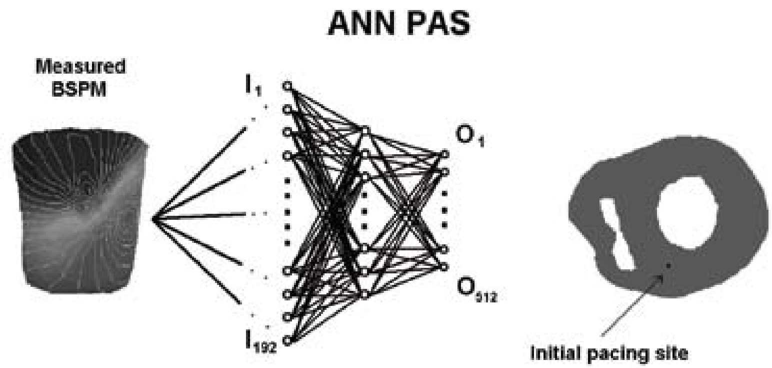
During the simulation, a parameter classification algorithm was applied to distinguish normal QRS complexes from abnormal ones, in order to determine the specific differences between the normal and abnormal parameter values. For normal cases, the detection ratio is practically 100%. The signals presented in Figure 3 were obtained via simulation using the initial parameter set for a normal and abnormal (bypass tract) situation.

Figure 4 presents the cell activation simulation that was effectuated and visualized after 200 ms of the rises of sinoatrial node excitation. In the right side of the image the activation time of each cell is showed. Neighbor slices are situated at 10 mm distance from each other, so totally a 5 centimeter wide ventricular tissue is visualized. Table 1 shows the efficiency of simulation for different cases. The evaluation of the simulated results was made by physicians. The performance was determined as the ratio of correct and total decisions.

DISCUSSION

Table 1 reveals that the 3D heart simulation (Szilágyi, Benyó, & Dávid Bey, 2003a, 2003b) succeeds in most cases, such as Wolff-Parkinson-White (WPW) syndrome, pre-excitations, and tissue activation model-

Figure 2. The ANN-based preliminary analyzer system used to estimate the pacing site from ECG



ling. Re-entry mechanisms and triggered event cases represent the weak points of the simulation model.

The application in practice of the model has several obstacles, which can be classified into the following groups:

- Effects of internal and external perturbations (such as environment, sympathetic and parasympathetic despondence).
- Lack of information on different elements of the model.
- Lack of technical background.

The Limitations of the Model

Several problems could be conceptualized, but the most important limitations are:

- The processes performed inside the cells are not well known, the behavior of the studied components cannot be determined with an acceptable precision.
- In critical cases, if a group of cells does not get the necessary food, it changes its behavior. A model created to simulate the normal behavior of the cell will not simulate it correctly in abnormal case.

Figure 3. The simulated ECG signal in normal and abnormal case (bypass tract)

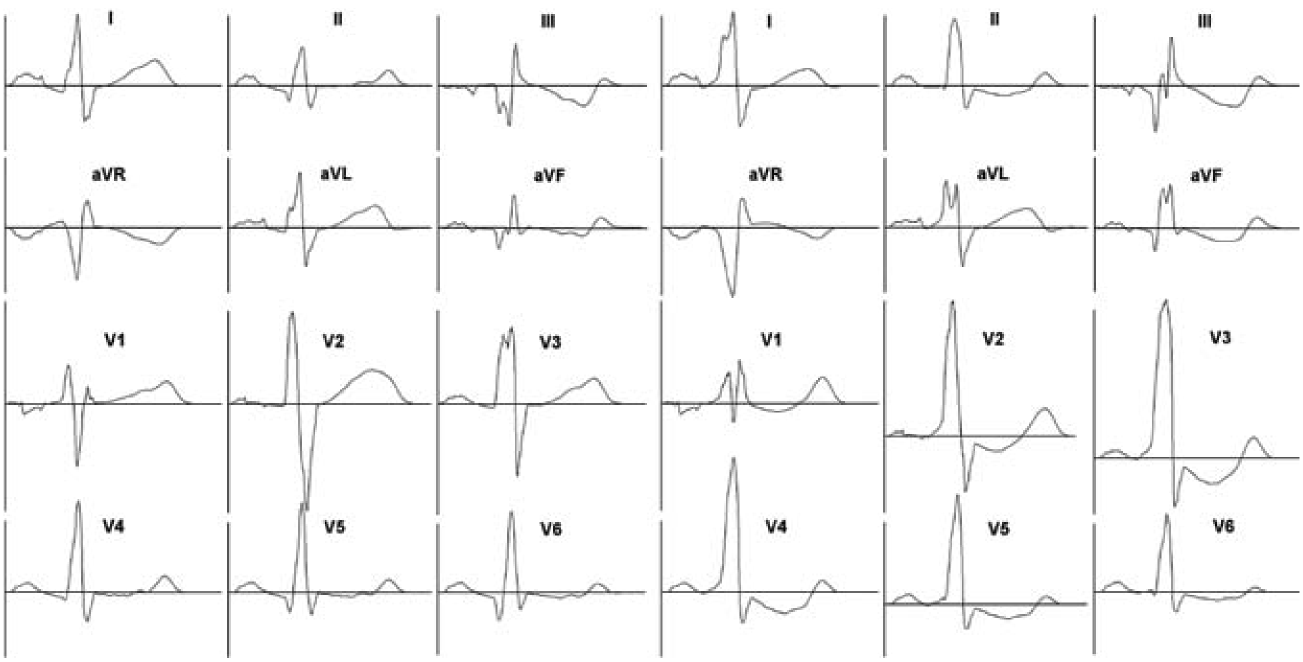


Figure 4. The simulated depolarization in normal and abnormal case (bypass tract)

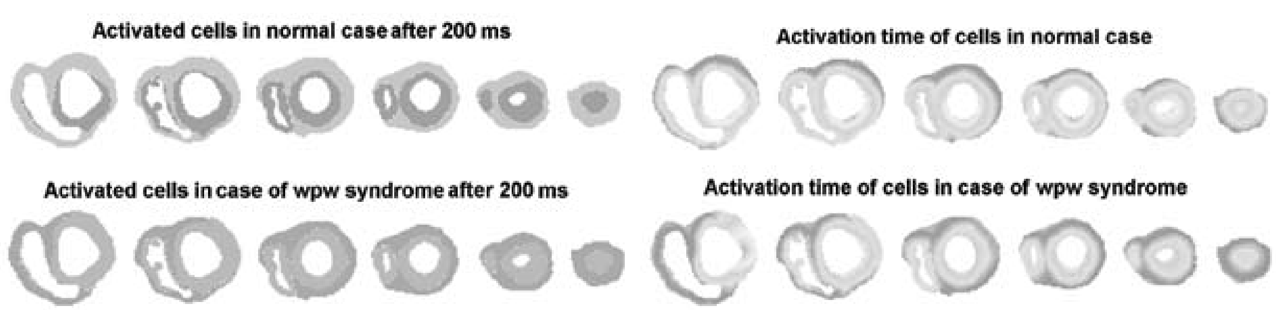


Table 1. Simulation performance in case of various pathological and normal situations

Pathological case	*Decision number*	*Failed decisions*	*Performance*
Normal	44	0	100.00%
Ectopic beat	21	0	100.00%
WPW syndrome	14	1	92.86%
Atrial flutter	22	1	95.45%
Atrial fibrillation	18	1	94.44%
Ventricular fibrillation	19	1	94.73%
Re-entry mechanisms	19	2	89.47%
Triggered activity	36	2	94.44%
Aberrant ventricular conduction	21	1	95.24%

- Because the structure of the heart differs from patient to patient, this structure is not known a priori and it has to be determined in real-time, based on the available information.
- Determining the torso's structure introduces the same problem. It is hard to determine the electrical conductivity and precise position of its elements.

Perturbation Phenomena

In case of human system identification, the most important disturbing phenomena are:

- It is known that respiration makes the heart change its shape and position. Although the motion of the heart can be tracked, it is not possible to determine from the ECG the amplitude of the motion.
- The continuous motion and displacement involves very hard problems. Because the motion has an effect on the behavior of all internal elements, the behavior of the heart will also be modified. The model has to follow the changes of the cell properties. For example: a resting man suddenly jumps out of the bed. The controlling mechanisms start their adjustment, the values of model parameters will change.
- Fever and respiration frequency can also cause alterations.
- External events (the patient senses something annoying or pleasant) change the dependence between the previously measured signals and the determined parameters. This is one of the causes why the perfect simulation of a human body is impossible.

Technical Background

At present, the performance of personal computers does not make possible the real-time determination of parameter values. The practical application is possible only in case of strongly parallel systems. The simplified model can be applied in real-time, but its efficiency is reduced because of the neglected parameters. The waveform of the simulated ECG in normal cases can be considered acceptable. The shape and duration of basic waves have realistic values. In case of abnormal cases the obtained waveform is not acceptable and more simulations are needed.

CONCLUSION

Regarding the fact that computerized ECG diagnostics refer to several medical and technical problems, at the moment it cannot be applied as a standalone system. The short-term solution is the application of fuzzy systems and systems based on multi-agents that, based on empirical information, make it possible to accomplish an adaptive advising system based on continuous transitions.

If a hybrid system (neuronal-fuzzy and model-based approach, simultaneously) is built, it may become possible to learn the model via the knowledge of the traditional advising system, which, after a suitable learning process, will be able to replace gradually the old system.

ACKNOWLEDGMENT

This research was supported in part by the Sapientia Institute for Research Programmes, Domus Hungarica Scientiarium et Artium, and the Communitas Foundation.

REFERENCES

Cuppen, J. J. M., & van Oosterom, A. (1984). Model studies with inversely calculated isochrones of ventricular depolarization. *IEEE Trans. Biomed. Eng., 31*, 652-659.

Guanglin, L., & Bin, H. (2001). Localization of the site of origin of cardiac activation by means of a heart-model-based electrocardiographic imaging approach. *IEEE Trans. Biomed. Eng., 48,* 660-669.

Lagerholm, M., Peterson, C., Braccini, G., Edenbrandt, L., & Sornmo, L. (2000). Clustering ECG complexes using hermite functions and self-organizing maps. *IEEE Trans. Biomed. Eng., 47,* 838-847.

MacLeod, R.S., & Brooks, D.H. (1998). Recent progress in inverse problems in electrocardiology. *IEEE EMBS Magazine, 17(1),* 73-83.

Minami, K., Nakajima, H., & Yoyoshima, T. (1999). Real time discrimination of the ventricular tachyarrhythmia with Fourier-transform neural network. *IEEE Trans. Biomed. Eng., 46,* 179-185.

Mirvis, D. M. (1988). Validation of body surface electrocardiographic mapping. In D. M. Mirvis (Ed.), *Body surface electrocardiographic mapping* (pp. 63-74). Boston, Dordrecht, London: Kluwer Academic Publ.

Osowski, S., & Hoai, L. T. (2001). ECG beat recognition using fuzzy hybrid neural network. *IEEE Trans. Biomed. Eng., 48,* 1265-1271.

Osowski, S., Hoai, L. T., & Markiewicz, T. (2004). Support vector machine-based expert system for reliable heartbeat recognition. *IEEE Trans. Biomed. Eng., 51*, 582-589.

Shahidi, A.V., Savard, P., & Nadeau, R. (1994). Forward and inverse problems of electrocardiography: Modeling and recovery of epicardial potentials in humans. *IEEE Trans. Biomed. Eng., 41,* 249-256.

Smola, A., & Cholkopf, B. (1998). A tutorial on support vector regression (NeuroColt Tech. rep. NV2-TR-1998-030). London: Royal Holloway College, Univ. London.

Szilágyi, S. M. (1998). Event recognition, separation and classification from ECG recordings. In *Proc. Ann. Int. Conf. IEEE EMBS*, Hong Kong (Vol. 20, pp. 236-239).

Szilágyi, S. M., Benyó, Z., & Dávid, L. (2003a). Heart model based ECG signal processing, In *Proc. 5th IFAC Symp. Modell. Contr. Biomed. Syst*., Melbourne (pp. 213-217).

Szilágyi, S. M., Benyó, Z., & Dávid, L. (2003b). WPW syndrome identification and classification using ECG analysis. In *Proc. World Congr. Med. Phys. Biomed. Eng*., Sydney.

Thaker, N.V., & Ferrero, J. M. (1998). Electrophysiologic models of heart cells and cell networks. *IEEE EMBS Magazine, 17(5)*, 73-83.

Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.

KEY TERMS

Body Surface Potential Map: A massively multichannel measurement of the ECG signal using several dozens or hundreds of electrodes that cover the whole torso

ECG Analysis: A collection of ECG signal processing methods aiming at the determination of medical parameters.

Heart Simulation: A simulation that focuses on the functional modeling of the heart.

Inverse Electrocardiography: A methodology that reconstructs the inner structure of the heart using multichannel ECG.

Search Space Reduction: A method for decreasing the problem complexity.

Support Vector Machine (SVM): A set of related supervised learning methods used for classification and regression.

Unified Neural Network: An advanced neural network whose criterion function includes the topology of the classified samples.

Wolff-Parkinson-White (WPW) Syndrome: A syndrome of pre-excitation of the ventricles of the heart due to an accessory pathway that causes an abnormal electrical communication from the atria to the ventricles.

Strategic Fit in the Healthcare IDS

Evelyn H. Thrasher
University of Massachusetts Dartmouth, USA

Terry A. Byrd
Auburn University, USA

INTRODUCTION

Interorganizational networks are defined as "clusters of organizations that make decisions jointly and integrate their efforts to produce a product or service" (Alter & Hage, 1993, p.2) and "advanced organizational structures perceived to improve efficiency, flexibility, and innovativeness and described as decoupled units developed because of rapid growth or knowledge and technology" (Schumaker, 2002). The healthcare integrated delivery system (IDS) is a distinct example of an interorganizational network. Defined as networks of healthcare organizations linked for the goals of clinical integration and an effective patient care continuum (Deluca & Enmark, 2002; Kilbridge, 1998; Young & McCarthy, 1999; Zucherman, Kaluzny, & Ricketts, 1995), IDSs may assume various organizational forms; namely, strategic alliances, contracted networks, or joint ventures, and may be comprised of multiple forms within a single network (Page, 2003). Also of interest is the distinction of the IDS as a lateral network of stakeholders, all directly serving the patient. This study uses the healthcare IDS to test a model of strategic fit and to examine differences in the nature and strength of the strategic fit to performance relationship across two distinct levels of IDS development.

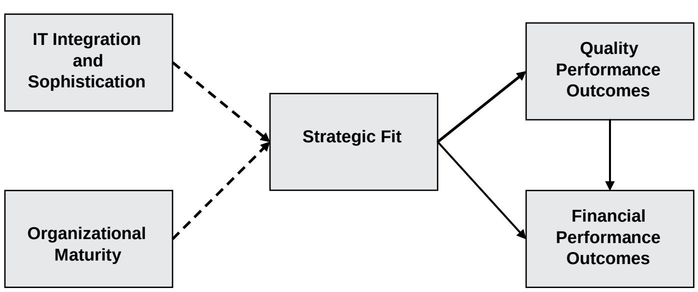
The conceptual model (Figure 1) is tested using data from HIMSS Analytics and the American Hospital Directory for 130 US IDSs. The IDSs are categorized into two levels of development, High Integration Aligned (HIA) and Low Integration Aligned (LIA), based on the alignment of IT integration/sophistication and organizational maturity. The relationship between strategic fit and IDS financial and quality performance is tested with comparisons made between the two groups.

BACKGROUND

Strategic Fit

Henderson and Venkatraman (1993), in the strategic alignment model, defined strategic fit as a process of adaptation in which organizational changes must be supported by complimentary IT resources and integration. Similarly, Chan and Huff (1993) defined strategic fit as "the fit between business strategy and IS strategy" (p. 345). We adapt these definitions to define strategic fit as the point of equilibrium at which the level of interorganizational network structure and maturity is properly aligned with a complementary level of IT integration and sophistication.

Figure 1. Conceptual model of strategic fit for the interorganizational network



Venkatraman (1994) developed the IT-Enabled Business Transformation Framework as an extension to the strategic alignment model, positing that IT is no longer simply an operational support resource, but rather a strategic tool with which to transform the firm's organizational structure and processes. He further proposed that IT's strategic role emerges more readily as firms establish strategic alliances and partnerships, as is the case with interorganizational networks. In advanced stages of organizational development, the need for and benefits from strategic fit should increase as interorganizational networks expand into more complex structures (Venkatraman, 1994). Yet little empirical evidence exists to support these suggestions. Much research around IT value and strategic fit has been conducted at the firm level (Bergeron, Raymond, & Rivard, 2001; Chan, Huff, Barclay, & Copeland, 1997), but research at the interorganizational network level is rare (Straub, Rai, & Klein, 2004).

Some researchers have examined the differences in firm performance between organizations at various levels of strategic fit. Zajac, Kraatz, and Bresser (2000) demonstrated significant positive links between strategic fit and ROA in savings and loan organizations that achieved advanced levels of strategic fit. These authors empirically supported the argument that organizations responding in a timely manner to needed strategy changes and achieving fit at this new level of development will realize greater benefits than those that do not.

Similarly, Bergeron, et al. (2001) found that small enterprises with a high level of strategic fit realized improved financial performance. These authors examined strategic fit through different lenses—profile deviation, moderation, matching, and other perspectives—to measure the performance impacts of fit. The results suggested that those organizations that pursue a highly strategic organizational and IT strategy tend to outperform organizations that fail to reach these higher levels. Thus, these past studies support the idea of a difference in the strategic fit to performance relationship, depending upon the achieved level of IT integration and organizational structure. More specifically, these studies suggest that performance improvements should be more apparent among highly integrated mature organizations that have achieved strategic fit.

Quality Outcomes

Quality outcomes are intangible, quality-related measures of organizational service and performance (Devaraj & Kohli, 2000; Li & Collier, 2000) and are of concern to IDSs due to the need for continual patient care quality improvement (Snyder & Paulson, 2002). Research suggests that patient-centered measures such as mortality rate, patient satisfaction, and average length of hospital stay (ALOS) are appropriate quality performance outcomes for the IDS (Devaraj & Kohli, 2000; Dowling, 1997; Smith & Swinehart, 2001); thus, ALOS is adopted for this study.

Initial formation and growth of IDSs led to reduced ALOS. For instance, Kim (2000) examined the impact of IDS formation on ALOS in the 1990s and found that ALOS was shorter for IDSs that had achieved a high degree of functional process integration across network participants. Those IDSs that had shortened ALOS had done so through the streamlining of patient care with expanded services and reductions in the time associated with administrative tasks. To the contrary, more recent studies indicate that ALOS reductions may have slowed or stalled (Page, 2003), suggesting that IDS formation alone is insufficient to ensure long-term quality improvements.

Regarding IT's role in healthcare performance improvements, Li and Collier (2000) investigated the impact of IT on performance through a survey of hospital administrators. The results indicated that IT had a significant positive impact on hospital quality and financial performance through IT's perceived ability to enhance accuracy, timeliness, and patient care effectiveness. Research suggests that increasing IT resource complexity and sophistication results in increasing pressure to pursue further business transformation for improved coordination and integration. Thus, as IDSs mature, IT's role seems to evolve from business process support to enabler of business transformation (Schumaker, 2002). Similarly, as IDSs expand, increasing differentiation of services often results, thereby forcing these networks to re-evaluate and improve both organizational and IT integration to maintain a streamlined patient care continuum (Schumaker, 2002).

Researchers suggest that strategic fit has a more significant impact on financial and quality performance among interorganizational networks with high IT integration and organizational maturity than among those with lower levels (Sabherwal & Chan, 2001; Smith & Swinehart, 2001). For instance, Sabherwal and Chan (2001) studied defenders, prospectors, and analyzers, and identified IT strategy profiles appropriate for various levels of organizational structure. These researchers found empirical support for a greater degree of performance improvement among more mature organizations with a focus on innovation and flexibility as opposed to immature organizations with an operational efficiency focus.

Extending the firm level findings and theories to the interorganizational network level, we propose the following hypothesis:

H1: The relationship between strategic fit and ALOS will be significantly more negative in HIA IDSs than in LIA IDSs.

Financial Outcomes

The current study uses operational cost as a measure of IDS financial performance to satisfy two conditions. First, by using cost measures, we can study both for-profit and not-for-profit IDSs, as both are concerned with reducing costs despite differences regarding profit-centered measures. Second, the healthcare industry is facing increasing pressure to reduce costs while continuing to improve patient care quality (Snyder & Paulson, 2002); thus, it is important to examine the impact of strategic fit on cost reduction. The formation of IDSs represents one attempt to control costs through the anticipated streamlining and improvement of the patient care continuum. Yet historical performance of IDSs has not supported this aim. Possibly due to the organizational complexity of newly formed healthcare networks, researchers suggest that organizational changes alone are insufficient to produce financial performance improvements for the IDS (Coddington & Moore, 2001).

In recent years, IDSs have begun to look closer at the benefits of IT integration in hopes of improving operational cost (Etchen & Boulton, 2000). Prior evidence suggests that IT may have a significant influence on the reduction of operational costs through improved efficiency and effectiveness (Barua, Kriebel, & Mukhopadhyay, 1995; Byrd, Thrasher, Lang, & Davidson, 2006; Coddington & Moore, 2001). For instance, a case study of 11 IDSs looked at steps taken to potentially reduce operational costs. In addition to the formation of the IDS and the streamlining of patient care within the IDS, all 11 stated that automation of clinical and administration processes through IT had resulted in significant operational cost reductions (Coddington & Moore, 2001).

Building upon these past results, Barua, et al. (1995) and Byrd, et al. (2006) also demonstrated support for the indirect cost benefits of IT associated with quality improvements. The premise behind these studies was that often the quality benefits of IT may be realized first and should, in turn, lead to financial performance improvements over time. Extending the firm-level evidence to the interorganizational network, we propose the following hypotheses:

H2: The relationship between strategic fit and operational cost will be significantly more negative in HIA IDSs than in LIA IDSs.

H3: The relationship between ALOS and operational cost will be significantly more positive in HIA IDSs than in LIA IDSs.

EMPIRICAL EVIDENCE OF IDS STRATEGIC FIT VALUE

Using 130 US IDSs, a strategic fit model was tested for the healthcare IDS. Profile matching (Chan et al., 1997) was used to categorize HIA IDSs and LIA IDSs, with 65 IDSs assigned to each group. Variables from the 2004 HIMSS Analytics Database were used to complete the categorization (Table 1).

Using standardized z scores and a six-point rating scale, each attribute was rated and averaged together to form an IT integration/sophistication rating and organizational maturity rating. Those scoring an overall rating of 4 or above were classified as HIA IDSs; those with an overall score of 3 or below were LIA IDSs. Table 2 presents descriptive statistics and ratings.

The measurement model was a second order model with IT integration/sophistication and organizational maturity as first-order factors forming strategic fit as a second-order factor. IT integration/sophistication and organizational maturity were measured reflectively using the factors in Table 1. The dependent variables, ALOS and operational cost, were measured using data from the 2004 American Hospital Directory.

Table 1. Variables for IDS classification

Construct	HIMSS Analytics Variables for the IDS
IT Integration and Sophistication	Number of enterprise applications in use Number of different types of enterprise applications (e.g., ERP, Clinical Decision Support, Case Mix Analysis) Number of available network nodes Number of personal computers in use Percentage of desktops with Internet access
Organizational Maturity	Number of facilities Number of different types of facilities (e.g., acute care, home healthcare, physician organization, insurer) Age (in years) of the IDS Number of hospital beds staffed Number of full-time employees Service population

Table 2. Descriptive statistics and IDS rating

Criteria	HIA IDS Rating N=65		LIA IDS Rating N=65	
	Mean	**SD**	**Mean**	**SD**
Enterprise Applications *(Mean=20.57, SD=23.26, Median=13)*	4.78	1.04	3.20	0.44
Types Enterprise Applications *(Mean=6.60, SD=0.70, Median=7)*	3.94	0.24	3.40	1.04
Available Network Nodes *(Mean=1744.35, SD=2460.84, Median=905)*	4.89	0.87	3.28	0.52
PCs in Use *(Mean=1272.42, SD=1809.48, Median=600)*	4.98	0.93	3.22	0.41
Percentage Personnel with Internet Access *(Mean=76.85%, SD=26%, Median=90%)*	3.75	0.66	3.35	0.89
Overall IT Integration and Sophistication	4.47	0.38	3.29	0.35
Facilities *(Mean=14.92, SD=17.24, Median=10)*	4.78	0.91	3.15	0.40
Diversity of Facilities/Services *(Mean=3.72, SD=1.02, Median=4)*	4.72	0.91	3.72	0.93
IDS Age *(Mean=56.79, SD=38.22, Median=52)*	2.94	1.04	3.66	1.05
Hospital Bed Capacity *(Mean=431.63, SD=490.38, Median=269)*	5.28	0.78	3.28	0.48
Full-Time Employees *(Mean=2256.95, SD=2903.53, Median=1300)*	5.22	0.80	3.26	0.44
Service Population *(Mean=1157848.28, SD=9172584.12, Median=270000)*	3.72	0.57	3.23	0.42
Overall Organizational Maturity	4.44	0.30	3.38	0.23
Overall Strategic Fit Rating	4.06	0.24	2.97	0.17

**Descriptives in italics*

Results

PLS was used for model analysis (Byrd et al., 2006; Sambamurthy & Chin, 1994), and T-tests were performed where needed to assess significance of differences. The results and measurement model are presented in Table 3 and Figure 2.

Discussion

Hypothesis 1 was supported, suggesting that HIA IDSs should see reduced ALOS, while ALOS may be increased in LIA IDSs. Researchers attribute increased efficiency, effectiveness, and knowledge sharing to high levels of IT integration and effective business processes. Perhaps these benefits contribute to improved ALOS in HIA IDSs. As researchers noted (Kim, 2000; Li & Collier, 2000), strategic fit with higher levels of IT and organizational integration may contribute to improvements in administrative processes and clinical care effectiveness, thereby resulting in shortened ALOS.

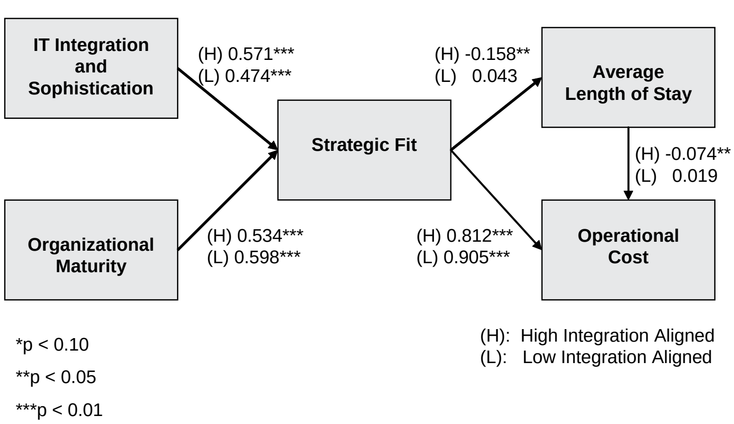
Hypothesis 2 was supported. Although positive for both groups, the strategic fit to operational cost relationship was significantly weaker for HIA IDSs, suggesting favorable progress is being made. History suggests that often the financial benefits of IT and strategic fit may not be readily apparent for a few years beyond the initial changes and investments. Thus, future research should consider including multiple years of operational cost data to better address the time lag issues of IT investments (Barua et al., 1995; Byrd et al., 2006).

Hypothesis 3 was not supported. HIA IDSs demonstrated a significant negative ALOS to cost relationship, while this relationship was insignificant and positive for LIA IDSs. One possible explanation is that as a patient's hospital stay is shortened, more frequent

Table 3. Hypotheses test results

	Path Coefficient		
Hypothesis	**HIA IDSs**	**LIA IDSs**	**t-Statistic for Difference**
H1: Strategic Fit to ALOS *(Supported)*	-0.158**	0.043	
H2: Strategic Fit to Operational Cost *(Supported)*	0.812***	0.905***	1.63*
H3: ALOS to Operational Cost *(Not Supported)*	-0.074**	0.019	
*p<.10. **p<.05. ***p<.01.			

Figure 2. Strategic fit model



patient turnovers may result in increased operational costs. Another possibility relates to the proposed indirect IT effect on financial performance through direct quality improvements (Barua et al., 1995; Byrd et al., 2006; Currie & Guah, 2006). Because a negative direct relationship between strategic fit and operational cost was not supported, the time lag effect of IT may also help to explain the lack of support for a positive relationship between ALOS and operational cost. As noted previously, using multiple years of cost data in future research might provide better insight into the effects of IT investment and strategic fit on operational cost, both directly and indirectly, through quality performance improvements (Barua et al., 1995; Byrd et al., 2006).

FUTURE TRENDS

Unfavorable historical performance has made IDS administrators hesitant regarding additional investments in IT, and some have considered dissolving IDS arrangements (Parker, Charns & Young, 2001). This study has demonstrated that those IDSs that have achieved strategic fit with high levels of IT integration/sophistication and organizational maturity should see improved quality and financial performance. An emphasis on IDS performance research is warranted and needed to further define the impact of strategic fit on a variety of healthcare performance measures.

IDS formation and IT integration are relatively young concepts in healthcare; therefore, it is probable that those who have reached the level of HIA have done so only in recent years. Thus, the time lag effect of IT investment suggests that direct effects on operational cost may not yet be fully realized. Further investigation of the proposed model is needed with longitudinal data and in varying contexts to further refine the model, to examine a broad spectrum of interorganizational arrangements, and to build a body of knowledge around IT's value for the IDS.

CONCLUSION

This study used the healthcare IDS to empirically test a model of strategic fit across two levels of interorganizational network development. The results suggest that the benefits of strategic fit differ across distinct levels of interorganizational development, thus further illuminating the complexities associated with taking IT research to the interorganizational network level of analysis. Even within a single industry, tremendous variation of interorganizational network form, structure, goals, and management increases the difficulty of empirical studies of IT value for these networks. Nevertheless, this level of analysis brings with it the potential for a rich body of knowledge around the issues of complexity, interorganizational structure, variance of scope and strategy, and other similar phenomena. Thus, the current study lays the groundwork for a research agenda centered around the complexities and nuances of the interorganizational network, an area of focus lacking in the IT literature.

REFERENCES

Alter, C., & Hage, J. (1993). *Organizations working together*. Newbury Park, CA: Sage Publications.

Barua, A., Kriebel, C.H., & Mukhopadhyay, T. (1995). Information technologies and business value: An analytic and empirical investigation. *Information Systems Research, 6*(1), 3–23.

Bergeron, F., Raymond, L., & Rivard, S. (2001). Fit in strategic information technology management research: An empirical comparison of perspectives. *Omega, 29*, 125–142.

Byrd, T.A., Thrasher, E.H., Lang, T., & Davidson, N. (2006). A process-oriented perspective of IS success: Examining the impact of IS on operational cost. *Omega, 34*(5), 448–460.

Chan, Y.E., & Huff, S.L. (1993). Investigating information systems strategic alignment. In *Proceedings of the International Conference on Information Systems*, 345–363.

Chan, Y.E., Huff, S.L., Barclay, D.W., & Copeland, D.G. (1997). Business strategic orientation, information systems strategic orientation, and strategic alignment. *Information Systems Research, 8*(2), 125–150.

Coddington, D.C., & Moore, K.D. (2001). The right strategy and perseverance can make an IDS profitable. *Healthcare Financial Management*, December, 35–39.

Currie, W. L., & Guah, M. W. (2006). IT-enabled healthcare delivery: The UK National Health Service. *Information Systems Management, Spring 2006,* 7-22.

Deluca, J.M., & Enmark, R. (2002). *The CEO's guide to healthcare information systems*. San Francisco: John Wiley & Sons, Inc.

Devaraj, J.M., & Kohli, R. (2000). Information technology payoff in the health-care industry: A longitudinal study. *Journal of Management Information Systems, 16*(4), 41–67.

Dowling, W.L. (1997). Strategic alliances as a structure for integrated delivery systems. In D.A. Conrad (Ed.), *Integrated delivery systems: Creation, management, and governance* (pp. 45–80). Chicago: Health Administration Press.

Etchen, L.L., & Boulton, J.L. (2000). Designing the IDS business portfolio. *Healthcare Financial Management*, January, 29–33.

Henderson, J.C., & Venkatraman, N. (1993). Strategic alignment: Leveraging information technology for transforming organizations. *IBM Systems Journal, 28*(2), 472–499.

Kilbridge, P.M. (1998). The role of information systems in IDS-physician relationships. *Healthcare Financial Management, 52*(6), 31–34.

Kim, Y.K. (2000). *An impact of integrated delivery systems on hospital financial and quality performance under managed care* [doctoral dissertation]. University of South Carolina.

Li, L.X., & Collier, D.A. (2000). The role of technology and quality on hospital financial performance. *International Journal of Service Industry Management, 11*(3), 202–224.

Page, S. (2003). Virtual healthcare organizations and the challenges of improving quality. *Healthcare Management Review, 28*(1), 79–92.

Parker, V.A., Charns, M.P., & Young, G.J. (2001). Clinical services lines in integrated delivery systems: An initial framework and exploration. *Journal of Healthcare Management, 46*(4), 261–275.

Sabherwal, R., & Chan, Y.E. (2001). Alignment between business and IS strategies: A study of prospectors, analyzers, and defenders. *Information Systems Research, 12*(1), 11–33.

Sambamurthy, V., & Chin, W. (1994). The effects of group attitudes toward GDSS designs on the decision-making performance of computer-supported groups. *Decision Sciences, 25*(2), 215–242.

Schumaker, A.M. (2002). Interorganizational networks: Using a theoretical model to predict effectiveness of rural healthcare delivery networks. *Journal of Health and Human Services Administration,* Winter, 372–406.

Smith, A.E., & Swinehart, K.D. (2001). Integrated systems design for customer focused healthcare performance measurement: A strategic service unit approach. *International Journal of Healthcare Quality Assurance, 14*(1), 21–28.

Snyder, K.D., & Paulson, P. (2002). Healthcare information systems: Analysis of healthcare software. *Hospital Topics: Research and Perspectives on Healthcare, 80*(4), 5–12.

Straub, D., Rai, A., & Klein, R. (2004). Measuring firm performance at the network level: A nomology of the business impact of digital supply networks. *Journal of Management Information Systems, 21*(1), 83–114.

Venkatraman, N. (1994). IT-enabled business transformation: From automation to business scope redefinition. *Sloan Management Review,* Winter, 73–87.

Young, D.W., & McCarthy, S.M. (1999). *Managing integrated delivery systems: A framework for action*. Chicago: Health Administration Press.

Zajac, E.J., Kraatz, M.S., & Bresser, R.K.F. (2000). Modeling the dynamics of strategic fit: A normative approach to strategic change. *Strategic Management Journal, 21*(4), 429–453.

Zucherman, H.S., Kaluzny, A.D., & Ricketts, T.C. (1995). Alliances in healthcare: What we know, what we think we know, and what we should know. *Healthcare Management Review, 20*(1), 54–64.

KEY TERMS

Average Length of Stay (ALOS): Number of days a patient must spend in the hospital, averaged across all lengths of patient stay for the hospital for one year.

Healthcare Integrated Delivery System: Network of healthcare organizations linked for the goals of clinical integration and an effective patient care continuum.

Interorganizational Network: A "cluster of organizations that makes decisions jointly and integrates their efforts to produce a product or service" (Alter & Hage, 1993, p.2).

Profile Matching: Profile matching assigns a score to both the IT integration/sophistication profile and the organizational maturity profile of the entity. If the two scores are equal or close within a very small predetermined range, the organization is deemed to have achieved strategic fit.

Strategic Fit: Point of equilibrium at which the level of interorganizational network structure and maturity is properly aligned with a complementary level of IT integration and sophistication.

Support Vector Machine Classification Applied on Weaning Trials Patients

B.F. Giraldo
Universitat Politècnica de Catalunya, Spain

A. Garde
CIBER de Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Spain

C. Arizmendi
Universitat Politècnica de Catalunya, Spain

R. Jané
Institut de Bioenginyeria de Catalunya (IBEC), Spain

I. Diaz
Hospital de la Santa Creu i Sant Pau, Spain

S. Benito
Hospital de la Santa Creu i Sant Pau, Spain

INTRODUCTION

The most common reason for instituting mechanical ventilation is to decrease a patient's work of breathing. Many attempts have been made to increase the effectiveness on the evaluation of the respiratory pattern by means of respiratory signal analysis. This work suggests a method of studying the lying differences in respiratory pattern variability between patients on weaning trials. The core of the proposed method is the use of support vector machines to classify patients into two groups, taking into account 35 features of each one, previously extracted from the respiratory flow. 146 patients from mechanical ventilation were studied: Group S of 79 patients with Successful trials, and Group F of 67 patients that Failed on the attempt to maintain spontaneous breathing and had to be reconnected. Applying a feature selection procedure based on the use of the support vector machine with leave-one-out cross-validation, it was obtained 86.67% of well classified patients into the Group S and 73.34% into Group F, using only eight of the 35 features. Therefore, support vector machines can be an interesting classification method in the study of the respiratory pattern variability.

Support vector machine (SVM) is a binary classifier based on a supervised statistical learning through examples (Cristianini, 2000). SVM introduced by Vapnik (1998) and studied by others research (Burges, 1998; Veropoulos, Cristianini, & Campbell, 1999) is a new and powerful learning methodology that can deal mainly with nonlinear classification and regression. SVM are based on the use of decision hyperplanes which determine decision boundaries. A decision hyperplane is one that separates the set of objects having different class memberships. The algorithm is simple enough to be analyzed mathematically, since it can be shown that correspond to a linear method in a high-dimensional feature space, nonlinearly related to input space (Hearst, 1998).

One of the most challenging problems in intensive care is the process of discontinuing mechanical ventilation. Critical-care clinicians must carefully weight the benefits of rapid liberation for mechanical ventilation against the risks of premature trials of spontaneous breathing and extubation. The need for accurate prediction applies to all phases of weaning, beginning with reductions in mechanical support, as patients are increasingly able to support their own breathing, followed by trials of unassisted breathing, which often precede extubation, and ending with extubation (Meade, Guyatt, Cook, & Griffith, 2001). When mechanical ventilation is discontinued, up to 25 percent of patients have respiratory distress severe enough to require reinstitution of ventilation support (Tobin, 2001).

The respiratory pattern describes the mechanical function of the pulmonary system and can be characterized by the following time series: inspiratory time (T_I), expiratory time (T_E), breath duration (T_{Tot}), tidal volume (V_T), fractional inspiratory time (T_I/T_{Tot}), mean inspiratory flow (V_T/T_I), and frequency-tidal volume ratio (f/V_T).

The aim of the present work is the analysis of the respiratory pattern variability in patients during weaning trials by means of support vector machines, in order to find differences between patients that had been capable to maintain spontaneous breathing, and patients that failed in that purpose.

The following method employs techniques from the area of signal processing and incorporates the doctor's knowledge, in order to achieve a satisfactory level of reliability so as to act as a decision support system in respiratory treatments. This research is an effort to exploit the capabilities of SVM as a classification method to identify and analyze the respiratory pattern variability.

ANALYZED DATA

Respiratory flow was measured in 146 patients on weaning trials from mechanical ventilation (WEANDB data base). These patients were recorded in the Departments of Intensive Care Medicine at Santa Creu i Sant Pau Hospital and Getafe Hospital, according to a protocol approved by the local ethic committees. Using clinical criteria based on the T-tube test, the patients were classified into two groups: Group S, 79 patients with Successful trials after 30 minutes and Group F, 67 patients that Failed in the purpose to maintain spontaneous breathing, and had to be reconnected after 30 minutes of weaning trials.

Respiratory flow was obtained using a pneumotachograph connected to an endotraqueal tube. The pneumotach consists in a Datex – Ohmeda monitor with a Validyne Model MP45-1-871 Feature-Reluctance Transducer. The signal was recorded at a sampling frequency of 250 Hz during 30 minutes.

From each recorded signal, the aforementioned time series were obtained: inspiratory time (T_I), expiratory time (T_E), breath duration (T_{Tot}), tidal volume (V_T), fractional inspiratory time (T_I/T_{Tot}), mean inspiratory flow (V_T/T_I), and frequency-tidal volume ratio (f/V_T)

METHODOLOGY

Data Preprocessing

Each one of the seven-time series was processed by moving a Running Window (RW), with a width range of 15, the best width ($p < 0.001$), consisting of several consecutive breath cycles. Five statistics were calculated for each window: the mean ($\overline{x}$), standard deviation (S), kurtosis (K), skewness (Sk), and interquartile range (IQR) of the value. In this way, 35 new time series were obtained for each patient.

Next, the data of each patient were analyzed independently by applying a *k*-means clustering algorithm, which automatically determines the number of clusters. In this study, for all patients, there was a main cluster containing the most part of the patterns (i.e., data points) with a considerable internal cohesion (low intra-cluster variance), corresponding to more than 96%, for each group. The patients were distributed in 80% for training process, and 20% for test process.

Leave-One-Out Cross-Validation

Leave-one-out cross-validation is a suitable technique to give a rough estimate of the performance of learning algorithm (SVM), and can be used to value the accuracy of the classifier (Ambroise & McLachlan, 2002). This external cross-validation should be undertaken to correct the error committed by choosing randomly the training and testing set. In this cross-validation process, each sample is removed in turn and the classifier is built using the remaining samples. The class of the removed sample is then predicted using the classifier created each case. This process is repeated for each sample in the original training set with which the classification accuracy is estimated.

The 35 features were ranked through a forward selection procedure, and beginning with the most relevant feature, the cross-validation error was estimated in turn adding the next relevant feature each time.

The best result was obtained with the next eight features: $S(T_E)$, $\overline{x}\ (T_{Tot})$, $\overline{x}\ (T_I)$, $\overline{x}\ (T_E)$, $S(T_{Tot})$, $IQR(T_{Tot})$, $\overline{x}\ (f/V_T)$, $\overline{x}\ (T_I/T_{Tot})$. Only these most relevant features were used in the final classification process, based on support vector machines.

Histogram Equalization

In order to achieve a reduction of the overlap between successful and unsuccessful patients (Group S and Group F), the variances of the features might be similar. Unfortunately, these variances cannot always be expected to be similar, so to solve this problem, an equalization of the histograms of the previously selected features as a nonaffine normalization process is proposed in this article (Hilger & Ney, 2001).

Histogram equalization or cumulative distribution function equalization is a general nonparametric method to make the cumulative distribution function (CDF) of some given data match to a reference distribution (Balchandran & Mammone, 1998).

The principle of this method is to find a nonlinear transformation to reduce the mismatch of two signals.

This transformation maps the distribution of a signal back to the distribution of the reference signal, and it is defined by means of the cumulative distribution functions of the signals in process.

In the equalization process, the objective is to deal with more reliable data by performing cumulative distribution function estimation by intervals equally spaced out. Each interval $x\ E[q_i, q_{i+1})$ is represented by $(x_i, F(x_i))$, that corresponds to the average of data and the cumulative distribution function obtained into the interval.

$$F(x_i) = \frac{k}{M} \qquad x_i = \frac{\sum_l x_l}{k_i}$$

where $x_i = x \in [q_i, q_{i+1})$, k is the number of data in the interval, $[q_0, q_{i+1})$, k_i is the number of data in the interval $[q_i, q_{i+1})$, and M is the total number of data.

The $F(x_i)$ defines the boundaries of the intervals in the CDF that will be equalized. These boundaries $[q'_i, q'_{i+1})$, limit the interval of values that fulfils the following expression $F(q_i) \leq F(y) < F(q_{i+1})$. All values of y that are into the interval $[q'_i, q'_{i+1})$, will be transformed to their corresponding x_i value.

The statistical matching technique applied in the current work matches the cumulative distribution function obtained from one feature, and the cumulative distribution function obtained from another feature, both evaluated over the training data.

The designed equalization takes, as a reference, the feature whereby the minor classification error is obtained on the leave-one-out cross-validation process which is the $S(T_E)$ feature.

Support Vector Machines

It is a learning system based on statistical learning theory, which can be applied to classification and regression problems.

SVMs deliver performance in applications such as text categorization, hand-written word recognition, image classification, biosequence analysis, and much more.

This learning method can be treated as a pattern recognition problem in which the features are seen as input patterns to be labeled as Group S patients or Group F patients obtained from the weaning off mechanical ventilation.

Given a linearly separable set of training vectors belonging to two separate classes, the aim is to find an optimal hyperplane that splits input data in two classes: 1, and -1 (the target values that correspond to the Group S and Group F patients labels respectively), maximizing the distance of the hyperplane to the nearest data of each class.

The separating hyperplane for linearly separable data is defined by $x \cdot w_0 + b = 0$, where w_0 is the normal to the hyperplane obtained as a linear combination of a subset of the training data. Data are then classified by computing the sign of $x \cdot w_0 + b = 0$. However, data normally are not separable; in this case, a linear boundary is inappropriate, a nonlinear decision function is needed, therefore an extension to nonlinear boundaries is achieved by using a specific function called kernel function. The kernel function maps the data of the input space into a higher dimensional space (feature space) by choosing a nonlinear transformation a priori. The SVM constructs an optimal hyperplane in this higher dimensional space (feature space), creating a nonlinear boundary in the input space. The mentioned hyperplane for nonlinearly separable data is defined by:

$$f(x) = \sum_{i=1}^{N} \alpha_i t_i\ K(x, x_i) + b$$

The coefficients α_i and b are determined by solving a large scale quadratic programming problem for which efficient algorithms exist that guarantee global optimum finding.

Where t_i are targets, and $\sum_{i=1}^{N} \alpha_i t_i = 0$

The vectors x_i are the support vectors, a linear combination of the support vectors determine the optimal separating hyperplane, and correspond to the closest points of each class. N is the number of support vectors and $0 \leq \alpha_i \leq C$. C is penalty parameter, which allows some flexibility in separating the categories. It controls the trade off between allowing training errors and forcing rigid margins, and creates a soft margin that permits some misclassifications. Increasing the value of C increases the cost of misclassifying points, and forces the creation of a more accurate model that may not generalize well. The goal is to find the minimum value of C with which a minimum error classification is obtained.

Furthermore, kernel functions must satisfy some constraints in order to be applicable (Mercer's conditions) (Burges, 1998).

The kernel function used in this work is a radial basis function (RBF), expressed as:

$$K(x_i, x_j) = \exp\left[-\frac{1}{2}\left(\frac{\|x_i - x_j\|}{\sigma}\right)^2\right]$$

where σ is a characteristic parameter of this kernel. Typically a method of clustering is first employed to select a subset of centers. An attractive feature of the SVM is that this selection is implicit, with each support vectors contributing one local Gaussian function, centered at that data point.

The best value of the parameter σ is estimated by an internal 10-fold cross-validation applied only over the training set.

A support vector machine is a robust classification technique capable to classify very overlapped data without taking into account the type of the data (Crawford, Miller, Shenoy, & Rao, 2005; Furey, Cristianini, Duffy, Bednarski, Shummer, & Haussler, 2000; Georgoulas, Stylios, & Groumpous, 2006).

RESULTS

The data set was composed of 146 patients (79 patients of Group S, and 67 patients of Group F), of which the 80% (116 patients) formed the training set and the remaining 20% (30 patients) the testing set.

SVM classifier was learned from the training data. Using the whole set of features (35), the average of correct classification was 66.67%. Firstly, an internal M-fold cross-validation, shows that the best value of the parameter σ used in the kernel function is σ = 0,5. Secondly, an external leave-one-out cross-validation was applied to get the classification error obtained with each feature, with an SVM trained using one unique feature. This error estimation showed the contribution of each feature to the performance of the classifier, thus it was determined which of the features were the most important, and had to be taken into account in the final classification to get the best result.

The classification performance was exhaustively examined by means of a forward selection process, and the best result was obtained with the following features: $S(T_E)$, $\bar{x}(T_{Tot})$, $\bar{x}(T_I)$, $\bar{x}(T_E)$, $S(T_{Tot})$, $IQR(T_{Tot})$, $\bar{x}(f/V_T)$, $\bar{x}(T_I/T_{Tot})$.

Once the features were selected, the remaining noisy features are removed; the computational cost reduced, and so did the classification error.

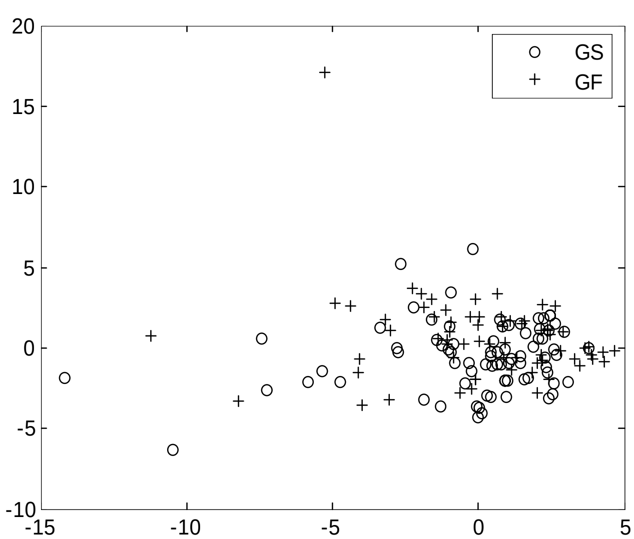
Histogram equalization technique was applied to the selected features with the aim of making their accumulation function be similar to the accumulation function of the reference feature that had given the best classification.

The classification percentage obtained after the process developed before is:

- Global percentage of well-classified patients: 80%
- Percentage of well classified Successful patient: 86.67%
- Percentage of well classified Unsuccessful patient: 73.34%
- The previous results, led to a sensitivity percentage of S=79%, specificity of E=82%, and value predictive positive of VPP=86%.

A grid-search was done to find the optimum penalty parameter *C*. As it was checked, the minimum *C*, which had given the best classification error accuracy, was

Figure 1. Successful (GS) and unsuccessful (GF) patients before the classification; orthogonal projections (Principal Component Analysis [PCA]) is used to represent hyperspace data on a plane before the classification



selected ($C = 15$), since the cost and the generalization error would be dropped.

Orthogonal projections (Principal Component Analysis [PCA]) were used as an attempt to visualize the hyperspace data variability on a plane before (Figure 1) and after the classification (Wang, Wu, & Liang, 2004).

Figure 2 and Figure 3 show the final classification of the training set and test set respectively.

CONCLUSION

Support vector machines is a robust classification technique capable to classify very overlapped data without taking into account the type of the data, leave one out cross-validation techniques allow searching the best subset of input features and histogram equalization maps the distribution of the selected features back to the distribution of the feature that gives the best classification result.

An innovative method based on support vector machines has been applied in order to analyze the respiratory pattern variability in patients during weaning trials, in order to find differences between patients capable of maintaining spontaneous breathing and patients that fail in that purpose.

The SVM trained with the whole set of 35 features achieved a well classification percentage of 66.67%. After the search of the most suitable features and the application of the histogram equalization method taking as reference the $S(T_E)$ feature, 80% was obtained using only eight of the 35 features. Although the groups are balanced, a greater facility of learning and generalization is observed on the part of the Group S with 86.67% of correctly classified percentage compared with Group S with 73.34%.

According to this classification, the following values of sensitivity S=79%, specificity E=82%, and value predictive positive VPP=86% were obtained.

These results permit to consider SVM analysis as a promising methodology to study the respiratory pattern variability in patients on weaning trials.

ACKNOWLEDGMENT

The authors would like to thank to Dr. A. Ballestero of the *Hospital Universitario de Getafe*, Getafe, Spain, by his collaboration in the signal database acquisition.

Figure 2. Training set of successful (GS) and unsuccessful (GF) patients after the classification; note that appropriate feature selection and histogram equalization process obtained with the training set are previous to the final classification

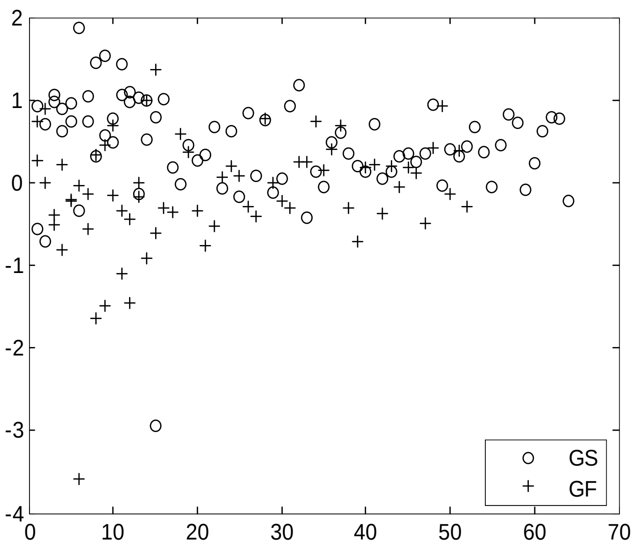
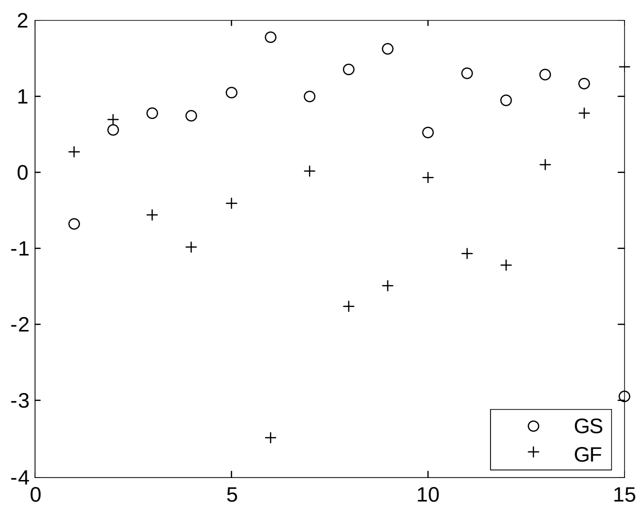


Figure 3. Test Successful (GS) and unsuccessful (GF) patients after the classification; note that appropriate variable selection and histogram equalization processes, obtained with the training set are previous to the final classification



This work was supported in part by *Ministerio de Educación y Ciencia* and *FEDER*, under grants TEC2007-63637, TEC2007-68076-C02-01 from the Spanish government.

REFERENCES

Ambroise, C., & McLachlan, G.J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *PNAS, 99*, 10.

Balchandran, R., & Mammone, R. (1998). Nonparametric estimation and correction of nonlinear distortion in speech systems. In *Proceedings of IEEE Int. Conf. Acoust. Speech Signal Proc*.

Burges, C.J.C. (1998). A tutorial on support vector machines for pattern recognition, data mining and knowledge discovery. *Kluwer Academic Publishers, 2*, 121–167.

Crawford, B., Miller, K., Shenoy, P., & Rao, R. (2005). Real-time classification of electromyographic signals for robotic control. In *Proceedings of AAI* (pp. 523–528).

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines*. Cambridge University Press.

Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Shummer, M., & Haussler, D. (2000). Support vector machine classification of cancer tissue samples using microarray expression data. *Bioinformatics, 6*(10).

Georgoulas, G., Stylios, C.D., & Groumpous, P.P. (2006). Predicting the risk of the metabolic acidosis for newborns based on fetal heart rate signal classification using support vector machines. *IEEE Transactions on biomedical engineering*, *53*(5).

Hearst, M. (1998). Trends and controversies: Support vector machines. *IEEE Intelligent Systems, 13*(4), 18–28.

Hilger, F., & Ney, H. (2001). Quantile based histogram equalization for noise robust speech recognition. In *Proceedings of EUROSPEEC, 2001* (pp. 1135–1138).

Meade, M., Guyatt, G., Cook, D., & Griffith, L. (2001). Predicting success in weaning from mechanical ventilation. *Chest*, *120*, 6.

Tobin, M.J. (2001). Advances in mechanical ventilation. *N. Engl. J. Med.,* 344(26), 1986–1996.

Vapnik, V.N. (1998). *Statistical learning theory.* New York: John Wiley & Sons.

Veropoulos, K., Cristianini, N., & Campbell, C. (1999). The application of support vector machines to medical decision support: A case study. In *Advanced Course in Artificial Intelligence (ACAI'99)*.

Wang, C., Wu, C., & Liang, Y. (2004). Medicine composition analysis based on PCA and SVM. *College of Computer Science and Technology, Jilin University, Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, Changchun 130012.* China.

KEY TERMS

Mechanical Ventilation: Breathing via machines.

Respiratory Pattern Variability: Seeing the breathing pattern.

Spontaneous Breathing: Natural breathing.

Support Vector Machine (SVM): A binary classifier based on a supervised statistical learning through examples, and is a new and powerful learning methodology that can deal mainly with nonlinear classification and regression.

Support Vector Machines in Neuroscience

Onur Seref
University of Florida, USA

O. Erhun Kundakcioglu
University of Florida, USA

Michael Bewernitz
University of Florida, USA

INTRODUCTION

In a typical *binary classification* problem, each pattern vector $\mathbf{x}_i \in \Re^d, i = 1,...,n$ belongs to one of two classes S^+ and S^-. A vector is given the label $y_i = 1$ or $y_i = -1$ if $\mathbf{x}_i \in S^+$ or $\mathbf{x}_i \in S^-$, respectively. The set of pattern vectors and their corresponding labels constitute the *training set*. The classification problem consists of determining which class new pattern vectors from the test set belong to. SVMs solve this problem by finding a hyperplane ($\mathbf{w}$, b) that separates the two classes in the training set from each other with the maximum margin.

The underlying optimization problem for the maximal margin classifier is only feasible if the two classes of pattern vectors are linearly separable. However, most of the real life classification problems are not linearly separable. Nevertheless, the maximal margin classifier encompasses the fundamental methods used in standard SVM classifiers. The solution to the optimization problem in the maximal margin classifier minimizes the bound on the generalization error (Vapnik, 1998). The basic premise of this method lies in the minimization of a convex optimization problem with linear inequality constraints, which can be solved efficiently by many alternative methods (Bennett & Campbell, 2000).

A hyperplane can be represented by $\langle \mathbf{w} \cdot \mathbf{x} \rangle + b = 0$ where $\mathbf{w}$ is the normal vector and b is the offset parameter. There is an inherent degree of freedom in specifying a hyperplane as $(\lambda \mathbf{w}, b\ \lambda)$. A *canonical hyperplane* is the one from which the closest pattern vector has a functional distance of 1, that is, $\min_{i=1,...,m} \left| \langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b \right| = 1$.

Consider two pattern vectors $\mathbf{x}^+$ and $\mathbf{x}^-$ belonging to classes S^+ and S^-, respectively. Assuming these pattern vectors are the closest to a canonical hyperplane, such that $\langle \mathbf{w} \cdot \mathbf{x}^+ \rangle + b = 1$ and $\langle \mathbf{w} \cdot \mathbf{x}^- \rangle + b = -1$, it is easy to show that the margin between these pattern vectors and the hyperplane are both equal to $1/\|\mathbf{w}\|$. Maximizing this margin while satisfying the canonicality condition for the pattern vectors turns out to be the following optimization problem.

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \tag{1a}$$

subject to

$$y_i \left(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b \right) \geq 1 \quad \forall i = 1, \ldots, n \tag{1b}$$

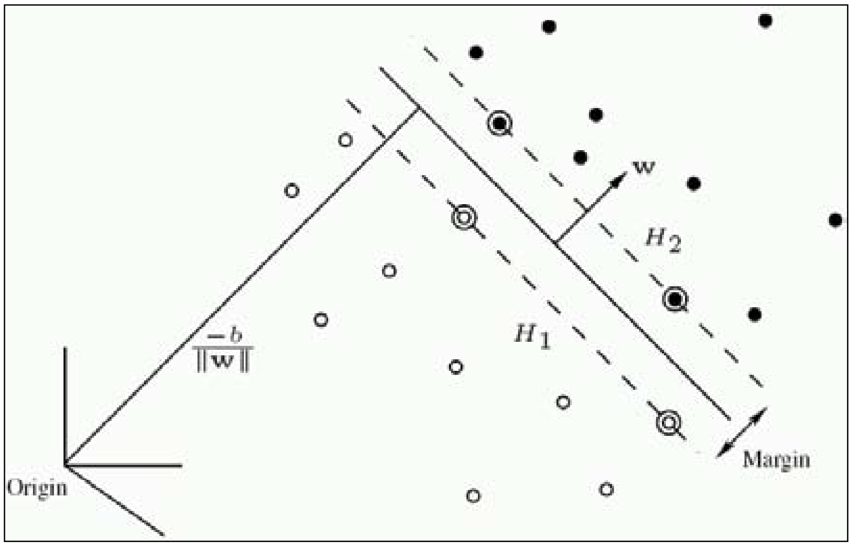
From the solution to (1), a new pattern vector $\mathbf{x}^*$ can be classified as positive if $\langle \mathbf{w} \cdot \mathbf{x}^* \rangle + b > 0$, and negative otherwise.

Most real life problems are composed of nonseparable data which is generally due to noise. In this case *slack variables* ξ_i are introduced for each pattern vector $\mathbf{x}_i$ in the training set. The slack variables allow misclassifications for each pattern vector with a penalty of $C/2$. In Figure 1-b, *soft margin classifier* is demonstrated that incurs penalty for misclassified pattern vectors. The maximum margin formulation can be augmented to soft margin formulation as follows:

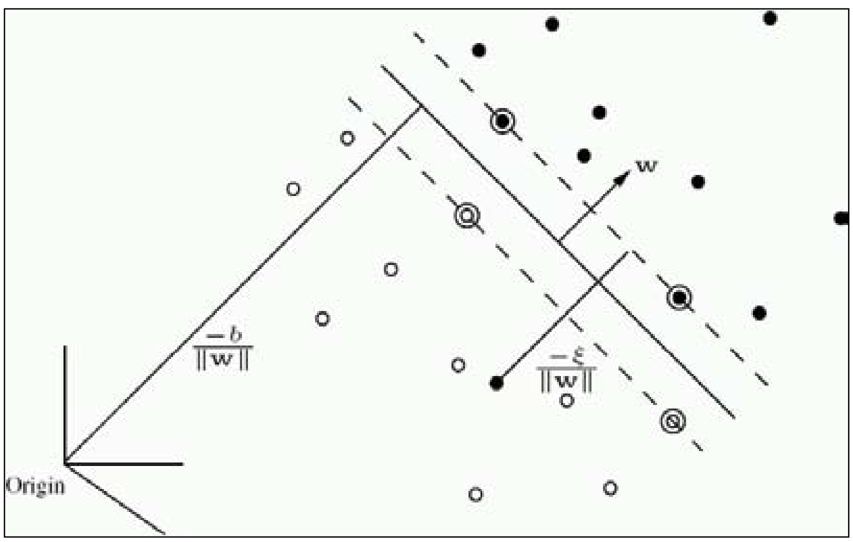
$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\sum_{i=1}^{n} \xi_i^2 \tag{2a}$$

subject to

Figure 1.



(a) Maximal margin classifier



(b) Soft margin classifier

$$y_i\left(\langle \mathbf{w}\cdot\mathbf{x}_i\rangle + b\right) \geq 1-\xi_i \quad \forall \mathrm{i}=1,\ldots,n \qquad \text{(2b)}$$

In (2), nonnegativity of the slack variables is implied since the solution cannot be optimal when $\xi_i < 0$ for any pattern vector.

The 2-norm of the slack variables is penalized in the objective of (2). An alternative formulation involves penalization of the 1-norm slack variables in the objective where nonnegativity of the slack variables is forced (Cristianini & Shawe-Taylor, 2000).

Dual formulations for both 1-norm and 2-norm SVMs can be obtained using the optimization theory. The significance of the dual formulations is that they do not involve inequality constraints, and they allow the *kernel trick* to be introduced for nonlinear classification. The standard method to obtain the dual formulation of the SVM problem consists of two parts. First the *Lagrangian function* of the primal problem is derived. This function provides a lower bound for the solution of the primal problem. Next, the Lagrangian function is differentiated with respect to the primal variables and stationarity is imposed. Equivalent expressions for each primal variable are substituted back in the Lagrangian function and added as constraints where appropriate. The dual problem is obtained by maximizing the resulting function with the new constraints. The dual problem is a concave maximization problem, which can also be solved efficiently. The Lagrangian function for the 2-norm SVM primal problem is given in Box 1.

Differentiating L with respect to the primal variables $\mathbf{w}$ and b, and assuming stationarity:

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{n} y_i \alpha_i \mathbf{x}_i = 0$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{n} y_i \alpha_i = 0$$

$$\frac{\partial L}{\partial \xi_i} = C\xi_i - \alpha_i = 0$$

Substituting the expressions back in the Lagrangian function, the following dual formulation is obtained which realizes the hyperplane $w^* = \sum_{i=1}^{n} y_i \alpha_i^* \mathbf{x}_i$ with geometric margin $\gamma = 1/\|\mathbf{w}^*\|$.

$$\max \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle - \frac{1}{2C}\sum_{i=1}^{n} \alpha_i^2 \tag{3a}$$

subject to

$$\sum_{i=1}^{n} y_i \alpha_i = 0 \tag{3b}$$

$$\alpha_i \geq 0 \quad \forall i = 1, \cdots, n \tag{3c}$$

Note that from Karush-Kuhn-Tucker complementarity conditions, the constraints in the primal problem are binding for those with the corresponding dual variable $\alpha_i^* > 0$. Therefore b^* can be calculated using:

$$b^* = \sum_{i:\alpha_i^*>0} y_i - \langle \mathbf{w}^* \cdot \mathbf{x}_i \rangle. \tag{4}$$

The decision rule $\operatorname{sgn}(f(\mathbf{x}))$ is equivalent to the hyperplane $f(\mathbf{x}) = \sum_{i=1}^{n} y_i \alpha_i^* \langle \mathbf{x} \cdot \mathbf{x}_i \rangle + b^*$.

Kernels are introduced in classification to provide enhanced similarity measures between pattern vectors. They basically transform the so-called *input space*, χ, in which the original pattern vectors reside, to a usually higher dimensional dot-product space H called the *feature space*, via a map $\Phi : \chi \to H$, such that $\mathbf{K}(\mathbf{x}_i \cdot \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) \rangle$. The kernel $\mathbf{K}$ is required to be positive semidefinite in order to define a dot product space and create a feature map. A positive semidefinite kernel is defined as a function on $\chi \times \chi$ for a nonempty set χ, which for all $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \chi$ gives rise to a positive semidefinite matrix $\mathbf{K}$ such that $\sum_{i,j} c_i c_j \mathbf{K}_{ij} \geq 0$ for all $c_i \in \Re$. In the literature it was shown that any algorithm that works on dot products can be kernelized through the *kernel trick* (Schölkopf & Smola, 2002). SVMs provide nonlinear classification by replacing the linear dot product $\langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle$ in the dual formulation with an appropriate nonlinear kernel $\mathbf{K}$.

One of the most commonly used kernels both theoretically and practically is the *Gaussian kernel* which is represented as:

$$\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma}} \tag{5}$$

where σ is referred to as the bandwidth. The smaller bandwidths are better in classifying intricate patterns, but worse in generalization.

APPLICATIONS IN NEUROSCIENCE

Although the introduction of SVMs is relatively recent, this classification technique is gaining rapid acceptance as a robust neuroscience and medical research

Box 1.

$$L(\mathbf{w}, b, \xi, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\sum_{i=1}^{n} \xi_i^2 - \sum_{i=1}^{n} \alpha_i \left[y_i \left(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b \right) - 1 + \xi_i \right]$$

tool. SVMs have been shown to be highly effective at detecting complex pattern changes in neuroimaging data. Often patterns that are difficult, time-consuming, or perhaps impossible to detect by the trained physician's eye can be detected rapidly and accurately using SVMs. In this manner, SVMs can assist doctors with the challenging task of performing rapid diagnosis of neurological diseases and disorders. Expediting diagnosis means expediting proper treatment that may mitigate or even prevent permanent damage to the central nervous system.

Many of the medical SVM applications focus on image processing of magnetic resonance imaging (MRI) data to detect structural alterations in the brain over time. Magnetic resonance spectroscopy (MRS) data has also been analyzed with SVMs for a similar purpose. Single photon emission computed tomography (SPECT) data can be used to detect blood flow pattern changes over time. Often these neurophysiologic measurement modalities can be utilized in an SVM classifier to detect the presence of a disease or disorder.

Numerous neuroscience studies have utilized SVMs to classify neural states. fMRI is a useful imaging modality for neural state classification due to its ability to track changes in blood oxygenation level dependent signal, which is correlated with blood flow. In addition, the electroencephalogram (EEG) is a highly useful measure for SVM neural state classification due to its ability to quantify the electrical activity in the brain (e.g., voltage difference between a region of interest and a reference region on the scalp or in the brain) with exceptionally high temporal resolution.

The remainder of this article will provide an overview of the state-of-the-art SVM applications to various neuroimaging modalities for the purposes of medical diagnosis, understanding the physiology of cognition, and classification of neural states.

Magnetic Resonance Imaging

There are a number of studies which introduce different methods for image *segmentation*. Segmentation is the process of partitioning a digital image into different sections in order to change the representation of the image. This new representation may involve certain characteristics in the image such as curves, edges, color, intensity, or texture. Segmented images are usually used to determine brain abnormalities using image classification, and it is shown that SVMs perform very well. Lee, Schmidt, Murtha, Bistritz, Sander, and Greiner (2005) used SVMs and a new SVM based method developed earlier called *support vector random fields* for segmenting brain tumors from MR images. Rinaldi et al. (2006) classified brain inflammation in *multiple sclerosis* patients based on the peripheral immune abnormalities from MR images using nonlinear SVMs. They determined that brain inflammation in patients with multiple sclerosis is associated with changes in subsets of peripheral lymphocytes. Thus, SVM classification helped detect a potential biomarker candidate for the prognosis of patients in the early stages of multiple sclerosis. Quddus, Fieguth, and Basir (2005) combined SVMs and *boosting* (Schapire, 2001), another machine learning method, to perform segmentation (via nonlinear classification) on white matter lesions in the MRI scans. Their composite classification method was shown to be faster and just as reliable as manual selection. In another study by Martinez-Ramon, Koltchinskii, Heileman, and Posse (2006), a similar approach was used to create segments of the brain with respect to their functions. Later, these segments were aggregated using boosting, which is used for multiclass SVM classification of an fMRI group study with interleaved motor, visual, auditory, and cognitive task design.

Other Imagery Types

Other imaging modalities such as ultrasound and MRS are also used in medical classification problems. Kotropoulos and Pitas (2003) used SVMs for segmentation of ultrasonic images acquired near lesions in order to differentiate between lesions and background tissue. The radial basis function SVMs outperformed the process of thresholding of L2 mean filtered images for various lesions under numerous recording conditions. Darbellay et al. (2004) used SVM and other classification techniques to detect solid or gaseous *emboli* by transcranial doppler (TCD) ultrasound. Since the leading cause of cerebral infarction is due to the extracranial atherosclerosis, rapid assessment of the physical characteristics of solid objects in the blood flow is important. Darbellay et al. (2004) demonstrated that SVMs could distinguish between solid and gaseous embolisms from ultrasonic measures of the blood stream. A medical diagnostic device based on this technology may be able to prevent brain damage by providing the means to expedite the diagnosis and treatment of embolisms. Devos et al. (2005) devised a system that can automatically discriminate brain tumors based on data from MRI and *MRSI*, which is a function of MR imaging that produces a spectroscopic profile of the scanned brain region. In this study, MR spectra from MRSI was used for comparison with linear and nonlinear *Least squares SVM* (LS-SVM) (Suykens & Vandewalle, 1999). A similar study was carried out (Devos et al., 2004) based on only MR spectra with *short echo time H-MRS*, using the same set of classifiers. Lukas et al. (2004), on the contrary, used *long echo H-MRS* signals to classify brain tumors into four classes: *meningiomas*, *gliobastomas*, *astrocytomas grade II,* and *metastases*, with a number of classifiers including standard SVM, and LS-SVM. The study demonstrated that kernel-based SVMs were able to detect tumors without utilizing dimensionality reduction and still produce accuracy comparable to linear discriminant analysis.

Automated tumor detection algorithms are a sought-after tool for assisting physicians to make more accurate and rapid detection of tumors. Support vector machine classifiers have contributed significantly in this area. Menze, Lichy, Bachert, Kelm, Schlemmer, and Hamprecht (2006) utilized SVM classification of MRI images to serve as an automated diagnostic tool for the detection of recurrent brain tumors. They reported that SVM among other methods was able to rule out lipid and lactate signals as being too unreliable, and that choline and N-acetylaspartate are the main sources of information (most important features).

Rapid diagnosis of stroke in patients is desirable as punctual treatment can reduce the chance of permanent brain damage. One potential method for rapidly diagnosing stroke is to examine the contents of a potential stroke patient's blood for biomarkers indicative of a stroke. Prados, Kalousis, Sanchez, Allard, Carrette, and Hilario (2004) utilized SVMs to help identify 14 potential biomarkers which could be used to distinguish the chemical profile of a control subject's blood from the chemical profile of an ischemic or hemorrhagic stroke patient. Surface enhanced laser desorption/ionization *mass spectometry* is used with SVMs for feature selection to find a small subset of potential biomarkers for early stroke diagnosis. Glotsos, Spyridonos, Cavouras, Ravazoula, Dadioti, and Nikiforidis (2005) and Glotsos, Tohka, Ravazoula, Cavouras, and Nikiforidis (2005) used digitized images of *biopsies* of *astrocytomas* to detect brain tumors with an *unsupervised SVM clustering* method in conjunction with a decision-tree based multiple class classification in order to classify tumors. Using morphological brain data and high-dimensional shape transformations, SVMs are also shown by Lao, Shen, Zue, Karacali, Resnick, and Davatzikos (2004) to classify between male and female brain and age differentiation for old adults. Although brain images are the primary sources for detecting brain abnormalities, electrical brain signals can also be used (Lehmann et al., 2007). They compared a number of classification methods including SVMs for the detection Alzheimer's disease from the EEG recordings and discovered that the SVMs performance was superior to other methods.

Feature Selection

Though it may seem counterintuitive, classification of large volumes of neurophysiologic data is not always very informative. A small portion of the data is usually sufficient for classification and the remaining data hinders the performance of the classifiers. *Feature selection* algorithms are developed to isolate such features which significantly contribute to the classification of the data. For very high resolution brain images or multiple-channel brain recordings, feature selection plays a very important role in classification of neural data. Fan, Shen, Gur, Gur, and Davatzikos (2007) introduced a method for classification of *schizophrenia* patients and healthy controls from brain structures whose volumetric features are extracted from processed MR images. The best set of such features are determined using an SVM-based feature selection algorithm, which in return significantly improved the classification performance. Yoon et al. (2007) extracted principal components derived from cortical thickness to differentiate between healthy controls and schizophrenic patients using SVMs as a diagnostic tool. Yushkevich, Dubb, Xie, Gur, Gur, and Gee (2005) investigated the effect of abnormal development and brain structure in patients with schizophrenia with respect to the morphological characteristics and age related changes. They used deformed brain templates of a variety of subject images and used SVMs for classification and feature selection to classify between pathological cases from the healthy controls. A similar study was carried out by Liu et al. (2004) for automated schizophrenia and Alzheimer's disease detection. Fung and Stoeckel (2007) also used an SVM feature selection algorithm applied to *SPECT* perfusion imaging to detect Alzheimer's disease. They used a *1-norm* linear SVM classifier, which is known to give sparse solutions, which in turn is used for feature selection.

Diabetes mellitus (DM) is a common disease in the industrialized countries and it is a prominent risk factor for ischemic cerebrovascular accidents. Diabetes alone is responsible for 7% of deaths in stroke patients. Diabetes mellitus often results in brain micro-blood flow disorders that may cause cerebral infarction. However, assessing the function of cerebral microvessels is difficult, since they are located within the bony skull. Kalatzis, Pappas, Piliouras, and Cavouras (2003) performed a study where SVM was applied to distinguish between blood flow data in patients with diabetes vs. control subjects using *SPECT* images from cerebral abnormalities. Li, Yang, Ye, and Geng (2006) used SVMs with *floating search method* to find relevant features for assessing the degree of malignancy in brain glioma from MRI findings and clinical data prior to operations. Li, Liu, and Cheng (2006) further developed a novel algorithm that combines bagging of SVMs with embedded feature selection for the individual observations and compared the new algorithm using publicly available datasets.

Brain Computer Interface

Brain Computer Interface (BCI) is an extremely challenging research area which could revolutionize available treatment options for patients with paralysis. The main function of a brain computer interface is to detect brain activity patterns associated with the intention to execute some motor task which the paralyzed subject cannot physically perform. The BCI responds to the detected intentions by activating an attached prosthetic device which can enact the task for the subject.

BCI devices typically utilize neurophysiologic measures which can be acquired for extended durations and with high time resolution. Though fMRI can provide highly useful information about the temporal hemodynamic response to changes in brain activity, BCI research using electrographic signals has the significant advantage of relatively simple acquisition hardware and superior time resolution. The state of the art of electrographic acquisition systems give them potential for the creation of implantable BCI devices. One of the numerous challenges that stand in the way of enacting such a device is the ability to classify the brain activity patterns in real time while maintaining acceptable sensitivity and specificity. Of the many methods utilized for this task, SVMs have demonstrated a large degree of success.

The majority of the applications focus on prosthetics for patients suffering from conditions such as amyotrophic lateral sclerosis (ALS), brainstem stroke, and brain injury. Guigue, Rakotomamonjy, and Canu (2006) developed a new graph based method to classify nonstationary signals, each with a discriminant waveform with random time and location. The graph based representation was used to define an inner product between graphs to be used with SVMs, which increased the accuracy of the BCI system.

Many studies have utilized the P300 evoked potential as an SVM input for classifying text which is read by the user (Thulasidas, Guan, and Wu, 2006). The P300 evoked potential is an event-related electrical potential which appears approximately 300 ms after an infrequent event is perceived. A P300 spelling device could provide the means for communication for the disabled individuals who would otherwise be unable to communicate with the world. This technique is frequently used to assess the performance of BCI related methods. Kaper and Ritter (2004) and Kaper, Meinicke, Grossekathoefer, Lingner, and Ritter (2004) used SVMs on EEG recordings from the P300 speller BCI paradigm to reach high rates of data transfer and generalization. In these studies the subjects were given a 6 by 6 matrix with flashing symbols and were instructed to attend to only one symbol to count how many times it appears. The SVM classifier was used to detect this P300 component in the EEG, and was shown to perform with high accuracy. Guan, Chen, and Lin (2005) used a similar mental speller paradigm with a target and non-target symbols moving from right to left in a small window on a computer screen and detected significant differences in the EEG using SVMs.

Since BCI systems need to be computationally efficient, a robust feature selection algorithm is required for prediction of the EEG channels that may be involved in the cognitive task of interest. Lal et al. (2004) investigated the feature selection and EEG signal classification problems with the SVM-based *Recursive Feature Elimination* and *Zero-Norm Optimization* methods.

Although nonlinear methods can provide better results, linear methods may be preferred wherever possible. However, complex cases still require efficient methods in BCI which can handle nonlinear classification such as SVMs (Müller, Anderson, & Birch, 2003). Garrett, Peterson, Anderson, and Thaut (2003) also used SVMs to classify EEG signals from a well known

EEG data set (Keirn & Aunon, 1990), which involve five different mental tasks, and showed that linear and nonlinear methods may perform similarly. Liang, Saratchandran, Huang, and Sundararajan (2006) used the Extreme Learning Machine (ELM) algorithm to classify EEG signals from the same data set and showed that ELM has similar performance to SVMs.

Some BCI systems are developed using non-human subjects. Rats are the most common subjects for this kind of research. A BCI system adapted for rats was developed by Hu, Si, Olson, and He (2005), who showed that SVM classifiers and principal component analysis combined with a Bayesian classifier may perform equally well for classification. They also showed SVM classification of neuronal spike trains allow identification of individual neurons associated with the decision making process. Jakuczun, Kublik, Wójcik, and Wróbel (2005) applied SVMs to classify habituated states from aroused states using evoked potentials from a single barrel column of the rat's somatosensory cortex. Olson, Si, Hu, and He (2005) used spike trains from rats to predict left and right hand commands in a binary paddle pressing task performed by rats.

Optical measurement methods have also demonstrated success in SVM BCI systems. Sitaram et al. (2007) used near infrared spectroscopy to detect oxygenation in the left hand vs. right hand motor imagery of human subjects from a 20-channel NIRS system. They used SVMs and hidden markov models to assess the feasibility of NIRS to be used as BCI and concluded that NIRS can be a significant alternative to EEG in BCI.

A study by Acir and Güzeliş (2005) investigated the utility of SVMs for identifying EEG sleep spindles, an EEG pattern found in stage 2 of sleep. The study demonstrated that radial basis SVMs detected EEG sleep spindles with high accuracy. This application of SVMs may be useful in an automated sleep staging algorithm.

Epilepsy is the condition of recurrent seizures. Over the past few decades, the area of seizure detection and seizure prediction using quantitative EEG analysis has drawn great interest. Chaovalitwongse, Prokopyev, and Pardalos (2006) developed a seizure prediction algorithm using SVMs which was able to successfully classify between EEG patterns associated with an interictal (normal) brain state and EEG patterns associated with a preictal (seizure prone) state. Such an algorithm could be developed to become the basis for a bedside or implantable seizure control device.

Cognitive Prediction

Prediction of cognitive states plays an important role in enhancing our understanding of the physiologic mechanisms underlying brain functions. Brouwer and van Ee (2007) used SVMs on functional fMRI data to predict the visual perceptual states from the retinotopic visual cortex and motion-sensitive areas in the brain. Cox and Savoy (2003) investigated visual presentation of various categories of objects. They used SVMs to classify the images based on similarity from predetermined regions of voxels (volume elements) over a short period of time. This method was shown to produce similar results using much less data than traditional fMRI data analysis, which requires numerous hours of data across many subjects. Pessoa and Padmala (2007) also used fMRI images to predict perceptual states. SVMs are used to detect near-threshold fear detection, and concluded that multiple regions of the brain are involved and that behavioral choice is distributed across these regions to help manage the emotional stimuli and prepare the appropriate response. Shoker, Sanei, and Sumich (2005) introduced a hybrid algorithm which combines SVMs with blind source separation (BSS) and directed transfer functions (DTF), which are used to extract features from EEG recordings to distinguish between left and right finger movements. Seref, Cifarelli, Kundakcioglu, Pardalos, and Ding (2007) used intracranial local field potential recordings from macaque monkeys and developed a selective SVM-based classification method in conjunction with SVM-based feature selection methods to detect categorical cognitive differences in visual stimuli based on single-trials from a visuomotor task.

Other Modeling Techniques

Faugeras et al. (2004) studied brain anatomy and modeled brain function from MR images. SVMs combined with methods from information theory are used in clustering of voxels in the statistical modeling of the fMRI signals. LaConte, Strother, Cherkassky, Anderson, and Hu (2005) used SVMs in block design fMRI and compared them to canonical variance analysis (CVA). Mouro-Miranda, Reynaud, McGlone, Calvert, and Brammer (2006) investigated the performance of

SVMs with time compression on single and multiple subjects, and showed that the time compression of the fMRI data improves the classification performance. In a similar study, Mouro-Miranda, Friston, and Brammer (2007) introduced time series embedding into the classification framework. In this work spatial and temporal information was combined to classify different brain states in cognitive tasks in patients and healthy control subjects. In a study by Wang, Schultz, Constable, and Staib (2003), a nonlinear framework for fMRI data analysis is introduced, which uses spatial and temporal information to perform *support vector regression* in order to find the spatio-temporal autocorrelations in the fMRI data. Finally, Parra, Spence, Gerson, and Sajda (2005) present an array of methods as "recipes" for linear analysis of EEG signals, among which performance of SVMs is compared with logistic regression.

CONCLUSION

This review presents the numerous applications in neuroscience which have benefited from the efficient and robust classification ability of SVMs. As computer capabilities improve and novel implementations are developed, new SVM based methods will reveal more about the brain and its complex functions.

REFERENCES

Acir, N., & Güzeliş, C. (2005). Automatic recognition of sleep spindles in EEG via radial basis support vector machine based on a modified feature selection algorithm. *Neural Computing and Applications, 14*(1), 56-65.

Bennett, K., & Campbell, C. (2000). Support vector machines: Hype or hallelujah? *SIGKDD Explorations, 2*(2), 1-13.

Brouwer, G. J., & van Ee, R. (2007). Visual cortex allows prediction of perceptual states during ambiguous structure-from-motion. *The Journal of Neuroscience, 27*(5), 1015-1023.

Chaovalitwongse, W. A., Prokopyev, O. A., & Pardalos, P. M. (2006). Electroencephalogram (EEG) time series classification: Applications in epilepsy. *Annals of Operations Research, 148*, 227-250.

Cox, D. D., & Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI) brain reading: Detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage, 19*, 261-270.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge, UK: Cambridge University Press.

Darbellay, G. A., Duff, R., Vesin, J. M., Despland, P. A., Droste, D. W., Molina, C., et al. (2004). Solid or gaseous circulating brain emboli: Are they separable by transcranial ultrasound? *Journal of Cerebral Blood Flow & Metabolism, 24*, 860-868.

Devos, A., Lukas, L., Suykens, J., Vanhamme, L., Tate, A., Howe, F., et al. (2004). Classification of brain tumours using short echo time 1H MR spectra. *Journal of Magnetic Resonance, 170*, 164-175.

Devos, A., Simonetti, A., Graaf, M. van der, Lukas, L., Suykens, J., Vanhamme, L., et al. (2005). The use of multivariate MR imaging intensities versus metabolic data from MR spectroscopic imaging for brain tumour classification. *Journal of Magnetic Resonance, 173*, 218-228.

Fan, Y., Shen, D., Gur, R. C., Gur, R. E., & Davatzikos, C. (2007). Compare: Classification of morphological patterns using adaptive regional elements. *IEEE Transactions on Medical Imaging, 26*(1), 93-105.

Faugeras, O., Adde, G., Charpiat, G., Chefd'Hotel, C., Clerc, M., Deneux, T., et al. (2004). Variational, geometric, and statistical methods for modeling brain anatomy and function. *NeuroImage, 23*, 46-55.

Fung, G., & Stoeckel, J. (2007). SVM feature selection for classification of spect images of Alzheimer's disease using spatial information. *Knowledge and Information Systems, 11*(2), 243-258.

Garrett, D., Peterson, D. A., Anderson, C. W., & Thaut, M. H. (2003). Comparison of linear, nonlinear, and feature selection methods for EEG signal classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 11*(2), 141-144.

Glotsos, D., Spyridonos, P., Cavouras, D., Ravazoula, P., Dadioti, P. A., & Nikiforidis, G. (2005). An image-analysis system based on support vector machines for automatic grade diagnosis of brain-tumour astrocytomas in clinical routine. *Medical Informatics and the Internet in Medicine*, *30*(3), 179-193.

Glotsos, D., Tohka, J., Ravazoula, P., Cavouras, D., & Nikiforidis, G. (2005). Automated diagnosis of brain tumours astrocytomas using probabilistic neural network clustering and support vector machines. *International Journal of Neural Systems*, *15*, 1-11.

Guan, J, Chen, Y., & Lin, J. (2005). Single-trial estimation of imitating-natural-reading evoked potentials in single-channel. In *Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference.*

Guigue, V., Rakotomamonjy, A., & Canu, S. (2006). Translation-invariant classification of non-stationary signals. *Neurocomputing*, *69*, 743-753.

Hu, J., Si, J., Olson, B. P., & He, J. (2005). Feature detection in motor cortical spikes by principal component analysis. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *13*(3), 256-262.

Jakuczun, W., Kublik, E., Wójcik, D. K., & Wróbel, A. (2005). Local classifiers for evoked potentials recorded from behaving rats. *Acta Neurobiol Exp*, *65*, 425-434.

Kalatzis, I., Pappas, D., Piliouras, N., & Cavouras, D. (2003). Support vector machines based analysis of brain spect images for determining cerebral abnormalities in asymptomatic diabetic patients. *Med. Inform.*, *28*(3), 221-230.

Kaper, M., Meinicke, P., Grossekathoefer, U., Lingner, T., & Ritter, H. (2004). BCI competition 2003-data set iib: Support vector machines for the p300 speller paradigm. *IEEE Transactions on Biomedical Engineering*, *51*(6), 1073-1076.

Kaper, M., & Ritter, H. (2004, Sep). Generalizing to new subjects in brain-computer interfacing. In *Proceedings of the 26th IEEE Embs Annual International Conference.* San Francisco, USA.

Keirn, Z. A., & Aunon, J. I. (1990). A new mode of communication between man and his surroundings. *IEEE Transactions on Biomedical Engineering*, *37*, 1209-1214.

Kotropoulos, C., & Pitas, I. (2003). Segmentation of ultrasonic images using support vector machines. *Pattern Recognition Letters*, *24*, 715-727.

LaConte, S., Strother, S., Cherkassky, V., Anderson, J., & Hu, X. (2005). Support vector machines for temporal classification of block design fMRI data. *NeuroImage*, *26*, 317-329.

Lal, T. N., Schrder, M., Hinterberger, T., Weston, J., Bogdan, M., Birbaumer, N., et al. (2004). Support vector channel selection in BCI. *IEEE Transactions on Biomedical Engineering*, *51*(6), 1003-1010.

Lao, Z., Shen, D., Xue, Z., Karacali, B., Resnick, S. M., & Davatzikos, C. (2004). Morphological classification of brains via high-dimensional shape transformations and machine learning methods. *NeuroImage*, *21*, 46-57.

Lee, C. H., Schmidt, M., Murtha, A., Bistritz, A., Sander, J., & Greiner, R. (2005). Segmenting brain tumors with conditional random fields and support vector machines. In *Computer Vision for Biomedical Image Applications* (pp. 469-478).

Lehmann, C., Koenig, T., Jelic, V., Prichep, L., John, R. E., Wahlund, L. O., et al. (2007). Application and comparison of classification algorithms for recognition of Alzheimer's disease in electrical brain activity (EEG). *Journal of NeuroscienceMethods*, *161*, 342-350.

Li, G. Z., Liu, T. Y., & Cheng, V. S. (2006). Classification of brain glioma by using SVMs bagging with feature selection. In *Data Mining for Biomedical Applications* (pp. 124-130).

Li, G. Z., Yang, J., Ye, C. Z., & Geng, D. Y. (2006). Degree prediction of malignancy in brain glioma using support vector machines. *Computers in Biology and Medicine*, *36*, 313-325.

Liang, N. Y., Saratchandran, P., Huang, G. B., & Sundararajan, N. (2006). Classification of mental tasks from EEG signals using extreme learning machine. *Int. J. Neural Syst.*, *16*(1), 29-38.

Liu, Y., Teverovskiy, L., Carmichael, O., Kikinis, R., Shenton, M., Carter, C., et al. (2004). Discriminative MR image feature analysis for automatic schizophrenia and Alzheimer's disease classification. In *Proceedings of the 7th International Conference on MedicalImage Computing and Computer Aided Intervention (MICCAI '04).*

Lukas, L., Devos, A., Suykens, J. A. K., Vanhamme, L., Howe, F. A., Majós, C., et al. (2004). Brain tumor classification based on long echo proton MRS signals. *Artificial Intelligence in Medicine*, *31*(1), 73-89.

Martinez-Ramon, M., Koltchinskii, V., Heileman, G. L., & Posse, S. (2006). fMRI pattern classification using neuroanatomically constrained boosting. *NeuroImage*, *31*, 1129-1141.

Menze, B. H., Lichy, M. P., Bachert, P., Kelm, B. M., Schlemmer, H. P., & Hamprecht, F. A. (2006). Optimal classification of long echo time in vivo magnetic resonance spectra in the detection of recurrent brain tumors. *NMR Biomed, 1* (5), 599-609.

Mouro-Miranda, J., Friston, K. J., & Brammer, M. (2007). Dynamic discrimination analysis: A spatial-temporal SVM. *NeuroImage, 36,* 88-99.

Mouro-Miranda, J., Reynaud, E., McGlone, F., Calvert, G., & Brammer, M. (2006). The impact of temporal compression and space selection on SVM analysis of single-subject and multi-subject fMRI data. *NeuroImage, 33*, 1055-1065.

Müller, K. R., Anderson, C. W., & Birch, G. E. (2003). Linear and nonlinear methods for brain-computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *11*(2), 165-169.

Olson, B. P., Si, J., Hu, J., & He, J. (2005). Closed-loop cortical control of direction using support vector machines. *IEEE Transactions on Neural Systems And Rehabilitation Engineering*, *13*(1), 72-80.

Parra, L. C., Spence, C. D., Gerson, A. D., & Sajda, P. (2005). Recipes for the linear analysis of EEG. *NeuroImage*, *28,* 326-341.

Pessoa, L., & Padmala, S. (2007). Decoding near-threshold perception of fear from distributed single-trial brain activation. *Cerebral Cortex*, *17,* 691-701.

Prados, J., Kalousis, A., Sanchez, J. C., Allard, L., Carrette, O., & Hilario, M. (2004). Mining mass spectra for diagnosis and biomarker discovery of cerebral accidents. *Proteomics*, *4,* 2320-2332.

Quddus, A., Fieguth, P., & Basir, O. (2005). Adaboost and support vector machines for white matter lesion segmentation in MR images. In *Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference.*

Rinaldi, L., Gallo, P., Ranzato, F., Luise, D., Colavito, D., Motta, M., et al. (2006). Longitudinal analysis of immune cell phenotypes in early stage multiple sclerosis: Distinctive patterns characterize MRI-active patients. *Brain*, *129*, 1993-2007(15).

Schapire, R. (2001). The boosting approach to machine learning: An overview. In *MSRI Workshop on Nonlinear Estimation and Classification*.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels*. Cambridge: MIT Press.

Seref, O., Cifarelli, C., Kundakcioglu, O. E., Pardalos, P. M., & Ding, M. (2007). Detecting categorical discrimination in a visuomotor task using selective support vector machines. In *Proc. of Biocomp '07,* (pp. 580-590).

Shoker, L., Sanei, S., & Sumich, A. (2005). Distinguishing between left and right finger movement from EEG using SVM. In *Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference.*

Sitaram, R., Zhang, H., Guan, C., Thulasidas, M., Hoshi, Y., Ishikawa, A., et al. (2007). Temporal classification of multichannel near-infrared spectroscopy signals of motor imagery for developing a brain-computer interface. *NeuroImage, 34*, 1416-1427.

Suykens, J., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters*, *9*, 293-300.

Thulasidas, M., Guan, C., & Wu, J. (2006). Robust classification of EEG signal for brain computer interface. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *14*(1), 24-29.

Vapnik, V. N. (1998). *Statistical learning theory*. New York: Wiley.

Wang, Y. M., Schultz, R. T., Constable, R. T., & Staib, L. H. (2003). Nonlinear estimation and modeling of fMRI data using spatio-temporal support vector regression. *Lecture Notes in Computer Science*, *2732*, 647-659.

Yoon, U., Lee, J. M., Im, K., Shin, Y. W., Cho, B. H., Kim, I. Y., et al. (2007). Pattern classification using principal components of cortical thickness and its discriminative pattern in schizophrenia. *NeuroImage*, *34*, 1405-1415.

Yushkevich, P., Dubb, A., Xie, Z., Gur, R., Gur, R., & Gee, J. (2005). Regional structural characterization of the brain of schizophrenia patients. *Academic Radiology, 12* (10), 1250-1261.

KEY TERMS

EEG: Electroencephalography is the measurement of the electrical activity of the brain from electrodes placed on the scalp or, in special cases, subdurally or in the cerebral cortex. The resulting traces are known as an electroencephalogram (EEG) and represent a summation of post-synaptic potentials from a large number of neurons.

fMRI: Functional magnetic resonance imaging is the use of MRI to measure the haemodynamic response related to neural activity in the brain or spinal cord of humans or other animals. It is one of the most recently developed forms of neuroimaging.

MRI: Magnetic resonance imaging, formerly referred to as magnetic resonance tomography (MRT), is a noninvasive method using nuclear magnetic resonance to render images of the inside of an object. It is primarily used in medical imaging to demonstrate pathological or other physiological alterations of living tissues. MRI also has uses outside of the medical field, such as detecting rock permeability to hydrocarbons and as a non-destructive testing method to characterize the quality of products such as produce and timber.

MRS: Magnetic resonance spectroscopy, also known as MRSI (MRS Imaging) and volume selective NMR spectroscopy, is a technique which combines the spatially-addressable nature of MRI with the spectroscopically-rich information obtainable from NMR. MR spectroscopy provides a wealth of chemical information on a region.

SPECT: Single photon emission computed tomography is a nuclear medicine tomographic imaging technique using gamma rays. It is very similar to conventional nuclear medicine planar imaging using a gamma camera. However, it is able to provide true 3D information. This information is typically presented as cross-sectional slices through the patient, but can be freely reformatted or manipulated as required.

SVM: Support vector machine is a set of related supervised learning methods used for classification and regression. SVMs belong to a family of generalized linear classifiers. They can also be considered a special case of Tikhonov regularization. A special property of SVMs is that they simultaneously minimize the empirical classification error and maximize the geometric margin; hence they are also known as maximum margin classifiers.

TCD: Transcranial doppler is a test that measures the velocity of blood flow through the brain's blood vessels. It is used to help in the diagnosis of emboli, stenosis, vasospasm from a subarachnoid hemorrhage (bleeding from a ruptured aneurysm), and other problems. It is often used in conjunction with other tests such as MRI, MRA, carotid duplex ultrasound, and CT scans.

Technology Adoption, Expectancy Value and Prediction Models

María Teresa Martín
Rey Juan Carlos University, Spain

María Victoria Román
Rey Juan Carlos University, Spain

Manuel Recio
University of Almería, Spain

INTRODUCTION

During the last few decades, various theoretical developments have been carried out with a view to describing the characteristic and distinct behavioral process that lies under any adoption of technological services and products. These developments are based mainly on the Social Psychology approach.

There are three extensive theories within the field of Social Psychology whose ultimate purpose has been to define the internal psychological factors that explain human behavior: the expectancy-value theory, the cognitive dissonance theory, and the self-perception theory. While the expectancy-value theory has been widely used in the research of adoption and usage of information systems, the other two theories have been less recognized.

Of all expectancy-value theory models, we should draw our attention to the reasoned action model (Azjen & Fishbein, 1980), because it underlies many of the studies on usage of technology. The planned behavior model (Azjen, 1985, 1991) represents a reformulation of the reasoned action model, justified by the existence of conducts that, albeit in part, a person cannot voluntarily keep under control. A rough description of both models is presented in this chapter, inasmuch as they served as a basis for the construction of the technology acceptance model (Davis, 1989; Davis, Bagozzi & Warshaw, 1989), known as one of the main models for the technology readiness concept. The technology acceptance model seems to possess a similar or even better explicating power than its predecessors (Davis et al., 1989; Mathieson, 1991; Taylor & Todd, 1995a; Chau & Hu, 2002).

The TAM has also been successfully applied in the health care context in order to examine the suitability of the TAM in explaining physicians' decisions to accept telemedicine technology (Chau & Hu, 2002).

Recently, authors have attempted to apply current models of technology acceptance to the use of Internet technologies and e-commerce. In particular, several authors have selected the TAM in order to characterize individuals' perceptions and intentions toward the use of the Internet (Muthitacharoen & Palvia, 2002; Pavlou, 2003). Due to the technological characteristics of the Internet channel, the TAM should provide a good framework where the specific factors that influence the adoption and usage of e-health technologies can be integrated. Physicians are among the principal users of this technology and have profound influences on its success.

EXPECTANCY-VALUE MODELS

We now proceed with a brief examination of the main expectancy-value models that have served as a theoretical basis for models of use of technology prediction, so we can gain a much better insight of the latter.

The Reasoned Action Model

The reasoned action model (Azjen & Fishbein, 1980), whose ultimate aim is the prediction and understanding of human behavior determinants, argues that the latter is mostly under the subject's control and, as a consequence, can be forecast by observing the declaration individuals make of their intentions to carry out or not to carry out a particular type of conduct. Also, the behavioral intention directly depends on two factors: a personal one or an attitude toward behavior that represents the individual's either positive or negative evaluation of whether an action should or should not be done; and a social factor or subjective norm that is the personal awareness of social pressures that force us to take that course of action.

Attitudes can be explained by considering the information the individuals possess that shapes their beliefs in the consequences when a certain conduct becomes real. Attitudes do not depend solely on beliefs but also on the evaluation people make about each of those behavioral beliefs. Therefore, the drive of each belief (degree of certainty that upholds it) is boosted by the negative or positive intensity of the outcome of the conduct. The sum of all products results in an attitude.

Subjective norms can be indirectly assessed from two main components: first, the informative basis of individuals, which establishes normative beliefs, that is to say, beliefs about how other people or institutions (referees) judge how an individual should behave; second, the subject's motivation to observe the referees' directives. Like it happens with attitudes, each normative belief is boosted by the motivation to follow that opinion. The subjective norm is obtained from the sum of all these products.

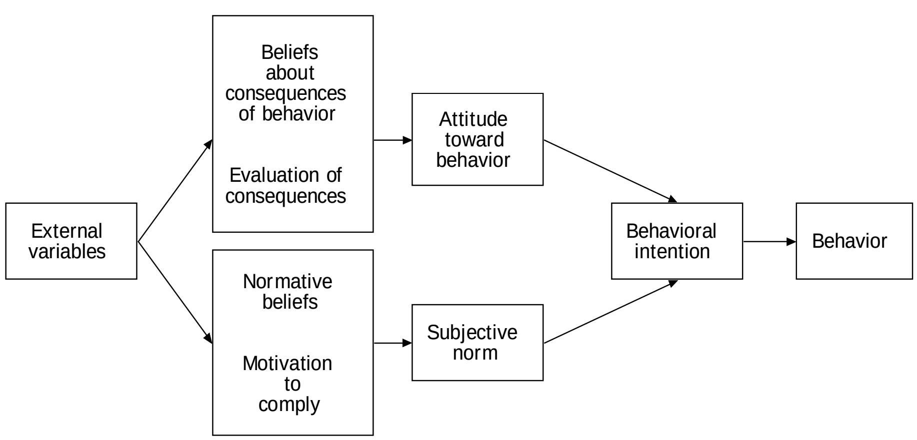
In this model, we can see a display of another series of external variables or stimulus conditions that affect relationships between intentions and behavior by means of their direct influence on beliefs in the consequences of conduct and in normative beliefs (see Figure 1).

Azjen and Madden (1986) introduced some nuances into the reasoned action model by pointing out that previous experience, attitudes, and subjective norm can influence one another; attitudes can influence behavior both indirectly through behavioral intention and also directly; subjective norm can only have an indirect effect on behavior; and previous experience can influence conduct indirectly as well as directly. The results of this investigation also show that the relation between intention and behavior is not as intense as it should be to make trustworthy predictions. Thus, "intentions only forecast behaviors not requiring either special abilities or skills, extraordinary opportunities, nor the other people's cooperation (Morales, Rebolloso & Moya, 1996, pp. 563–564).

The Planned Behavior Model

Azjen (1985, 1991) reformulated the aforementioned reasoned action model, thus originating the planned behavior model (see Figure 2). The difference between both models lies in the fact that the latter introduces the perceived behavioral control element to improve the behavioral intention forecast. The introduction of this intention-determining factor is justified by the existence of conducts, which at least in part cannot be kept under the voluntary control of the individuals. That is to say, irrespective of a likely favorable attitude toward a certain behavior and the strong pressure that individuals receive around them urging them to do it, their intention will

Figure 1. Reasoned action model



Source: Adapted from Ajzen and Fishbein (1980) also influence the subjective likeliness of achieving it or perceived behavioral control.

Perceived behavioral control depends on control beliefs, which decide whether individuals possess or do not possess the necessary skills or resources to carry out a certain conduct, and the suitable opportunities. Each control belief is boosted by the enhancing or inhibitory effect of the resource or opportunity dealt with. The final outcome after summing up all these products is the perceived behavioral control.

As the perceived behavioral control nears the control effectively exerted by individuals on their respective conducts, the relationship with the latter is direct.

PREDICTION MODELS OF USE OF TECHNOLOGY

The use of technology represents a generally accepted measure of the success of its widespread presence. Therefore, the aim of many investigations has been to determine which factors influence this kind of behavior.

In this section, we will look into the main theoretical models that aim to describe the characteristic behavioral process underlying the adoption of technological services and products. Through their analysis, we can detect the importance given to technology attitude as a key factor in predictions of use of technology.

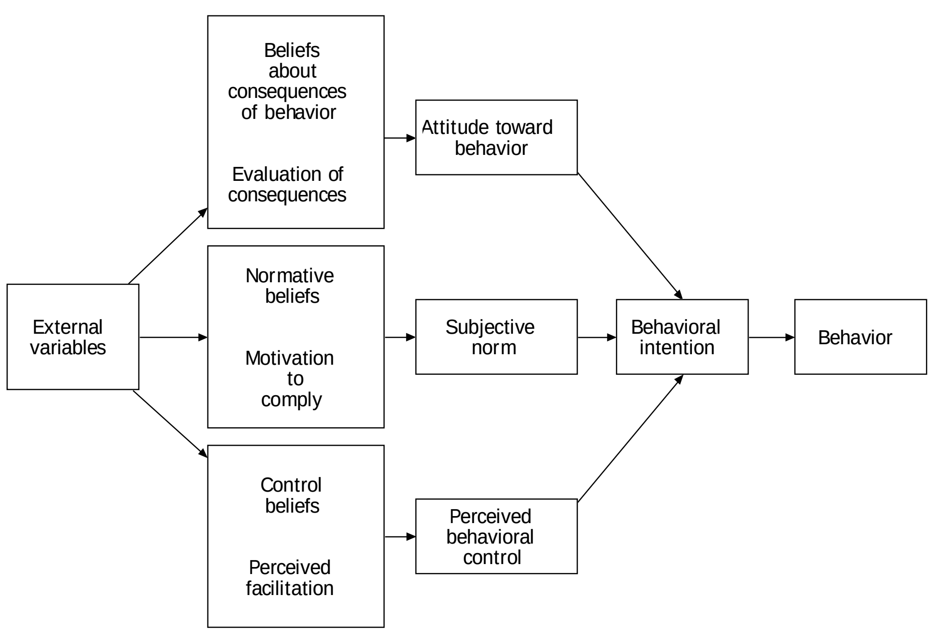
The review of the scientific literature has made clear that technology readiness is a multidimensional concept, although a consensus of opinion about the number and definition of these dimensions does not exist. In any case, such dimensions contain enhancing and inhibitory aspects of technology attitude.

With a view to completing the stage in which the domain of the studied concept is specified, after the exposition of each model, detailed aspects related to measure scales developed in various contexts and with different participants and technologies can be found. The examination of the various aspects of technology attitude (separately detected in the analyzed studies) will allow us to establish the suitability of its addition within the development of a measurement scale of the degree of technology readiness.

The Technology Acceptance Model

The technology acceptance model (Davis, 1989; Davis et al., 1989) is one of the most referred-to models for the explanation of adoption of information technologies. The acceptance of technology is firmly defined by

Figure 2. Planned behavior model



Source: Adapted from Ajzen (1985, 1991)

Gattiker (1984) as the psychological state of individuals with regard to their voluntary or desired use of a particular technology.

Based on the reasoned action model and the planned behavior model, the technology acceptance model put forward the idea that the behavior of adoption of a certain information system is determined by the intention of making use of it. In turn, such behavioral intention is determined by technology attitude, which is influenced by two factors: on the one hand, beliefs in perceived usefulness, and on the other hand, beliefs in the perceived ease of use of the analyzed technology. Perceived usefulness can be defined as the extent to which individuals think using the information system will improve their working productivity. The perceived ease of use can be defined as the extent to which an individual thinks the use of a system will not require any effort. Although component perceived usefulness has a direct effect on the behavioral intention, the perceived ease of use influences such intention only indirectly.

Furthermore, both types of factors are exposed to the influence of external variables. By operating on these variables, control can be exerted on the beliefs of the information system users, thus indirectly influencing the intentions of use and the effective utilization of technology. The described model is shown in Figure 3.

Technology Attitude Scales with a Theoretical Base on the Technology Acceptance Model

The technology acceptance model has been put into practice by making use of various types of technologies as a social object of interchange and considering various user profiles as a sample (see Table 1).

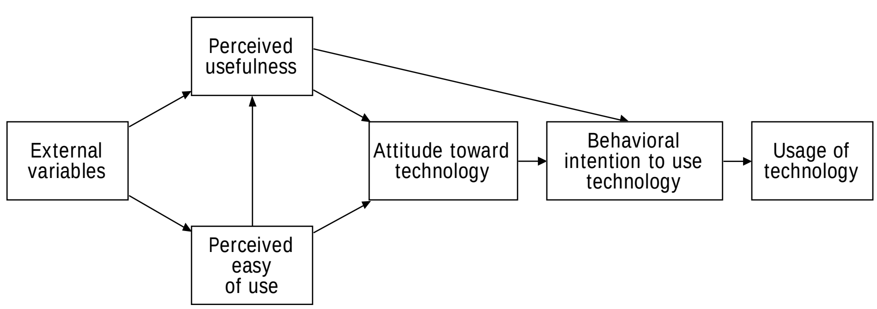
A review of various studies allowed us to note that among their authors there is a widespread inclination to remove the attitude component out of the model. They consider that behavioral intention is determined by two main components: on the one hand, beliefs in perceived usefulness, and on the other hand, beliefs in the perceived ease of use of the analyzed technology. In that way, the perceived usefulness component has a direct effect on the behavioral intention. The perceived ease of use influences such intention, indirectly and directly, through the perceived usefulness component.

Other authors, although considering attitude as a component of the model, identified no significant relation between attitude and behavioral intention.

The results obtained cannot lead us to erroneous conclusions about the dilemma of consideration or non-consideration of attitude as a key factor when predicting the use of technology. In the reviewed investigations, when attitude is taken into account as a component of the model, it is categorized as a one-dimensional concept. That is why simple scales have been applied to measure it, following the ones suggested by Ajzen and Fishbein (1980), introduced in the reasoned action model and planned behavior model. From our point of view, an attitude scale constructed like this lacks validity of content, insofar as it does not comprise all possible dimensions and contents of the analyzed concept.

The decision made by various authors to remove the attitude factor from the Technology Acceptance Model and, in its stead, to consider the direct effect of perceived usefulness and perceived ease of use factors on the behavioral intention, suggests to us that we should consider attitude as a multidimensional feature.

Figure 3. Technology acceptance model



Source: Adapted from Davis (1989) and Davis, Bagozzi and Warshaw (1989)

Therefore, attitude is not something global. It is made up of a combination of a series of variables or dimensions that determine attitude toward technology.

Once we established which dimensions constitute the domain of the attitude concept and made clear that it is a multidimensional one, the priority of the following step was to identify items in accordance with the conceptual dimensions as found. Next, we displayed in Table 2 the items included in the measurement subscale of perceived usefulness in the analyzed studies, whereas in Table 3 we placed the items of the measurement subscale of perceived ease of use.

As can be seen in the measurement of perceived usefulness, the use of the following items stands out:

- Using (information technology) increases my productivity.
- Using (information technology) increases my job performance.
- Using (information technology) enhances my effectiveness on the job.
- Overall, I find the (information technology) useful in my job.

Table 1. Application of technology acceptance model in various contexts

Software	User profile	Source
Text-editor	MBA students	Davis (1989)
Text-editor, e-mail	Managers and professionals	Davis (1993)
Spreadsheet	Students	Mathieson (1991)
Writeone, Chart-Master	MBA students	Davis, et al. (1989)
Voicemail system, customer dial-up system	(Not specified)	Subramanian (1994)
University computing, resource center, business school student	Students	Taylor & Todd (1995a)
University computing, resource center, business school student	Students	Taylor & Todd (1995b)
Configuration software	Salespersons	Keil, Beranek & Konsynski (1995)
E-mail	Graduate students	Szajna (1996)
C.A.S.E.	Information technology professionals	Chau (1996)
(Not specified)	Students	Venkatesh & Davis (1996)
Spreadsheet, database, word processor, graphics	Students	Jackson, Chow & Leitch (1997)
Personal computing	PC users	Igbaria, Zinatelli et al (1997)
Debugging tool	Students	Bajaj & Nidimolu (1998)
Configuration software	Salespersons	Gefen & Keil (1998)
Word processing spreadsheet graphics	Users of a Fortune 100 company	Agarwal & Prasad (1999)
Multifunctional workstation	Brokers and sales assistant of financial company	Lucas & Spitler (1999)
Microsoft Windows 3.1	Potential adopters and users in a corporation	Karahanna, Straub & Chervany (1999)
Telemedicine software	Physicians	Hu et al (1999)
Software maintenance tools	(Not specified)	Dishaw & Strong (1999)
Four different systems in four organizations	Floor supervisors, members of personal financial services	Venkatesh & Davis (2000)
Data and information retrieval	Professionals	Venkatesh & Morris (2000)
Telemedicine software	Physicians	Chau & Hu (2002)

With regard to the measurement of the perceived ease of use dimension, the items most frequently used are as follows:

- Learning to operate (information technology) is easy for me.
- I find it easy to get the (information technology) to do what I want to do.
- The (information technology) is rigid and inflexible to interact with.
- Overall, I find the (information technology) easy to use.

Limitations of the Investigations with a Theoretical Base on the Technology Acceptance Model

A review of the studies that used the technology acceptance model as a predictive model of the use of new technologies allowed us to identify some limitations of the research carried out. The most outstanding are the following:

- Starting with a selection of the sample to validate the model, we consider that the used samples comprise very specific population profiles. In

Table 2. Perceived usefulness questions

Perceived Usefulness	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Using (information technology) improves the quality of the work I do	●					●				
Using (information technology) gives me greater control over my work	●									
(Information technology) enables me to accomplish tasks more quickly	●					●	●	●		
(Information technology) supports critical aspects of my job	●									
Using (information technology) increases my productivity	●	●	●	●		●	●	●	●	●
Using (information technology) increases my job performance	●	●	●		●	●	●	●	●	●
Using (information technology) allows me to accomplish more work than would otherwise be possible	●									
Using (information technology) enhances my effectiveness on the job	●	●	●	●			●	●	●	●
Using (information technology) makes it easier to do my job	●			●		●	●	●		
Overall, I find the (information technology) useful in my job	●	●	●		●	●	●	●	●	●
(1) Davis (1989, 1993); (2) Davis et al. (1989); (3) Mathieson (1991); (4) Subramanian (1994); (5) Taylor & Todd (1995a, 1995b); (6) Keil et al. (1995); (7) Szajna (1996); (8) Chau (1996); (9) Jackson et al. (1997); (10) Igbaria et al. (1997)										

Perceived Usefulness	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
Using (information technology) improves the quality of the work I do		●	●		●					
Using (information technology) gives me greater control over my work			●							
(Information technology) enables me to accomplish tasks more quickly		●	●		●	●	●			●
(Information technology) supports critical aspects of my job										
Using (information technology) increases my productivity	●	●	●	●		●	●	●	●	●
Using (information technology) increases my job performance	●	●					●	●	●	
Using (information technology) allows me to accomplish more work than would otherwise be possible										
Using (information technology) enhances my effectiveness on the job	●		●	●	●	●	●	●	●	●
Using (information technology) makes it easier to do my job			●	●	●	●	●			●
Overall, I find the (information technology) useful in my job	●	●	●	●			●	●	●	
(11) Bajaj & Nidimolu (1998); (12) Gefen & Keil (1998); (13) Agarwal & Prasad (1997, 1999); (14) Lucas & Spitler (1999); (15) Karahanna et al. (1999); (16) Hu et al. (1999); (17) Dishaw & Strong (1999); (18) Venkatesh & Davis (1996, 2000); (19) Venkatesh & Morris (2000); (20) Chau & Hu (2002)										

Table 3. Perceived ease of use questions

Perceived Ease of Use	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
I find (information technology) cumbersome to use	•					•		•		
Learning to operate (information technology) is easy for me	•	•	•	•	•	•	•	•	•	
Interacting with the (information technology) is often frustrating	•					•				
I find it easy to get the (information technology) to do what I want to do	•	•	•			•	•		•	•
The (information technology) is rigid and inflexible to interact with	•			•			•	•		•
It is easy for me to remember how to perform tasks using the (information technology)	•									
Interacting with the (information technology) requires a lot of mental effort	•					•				•
My interaction with the (information technology) is clear and understandable	•					•				
I find it takes a lot of effort to become skillful at using the (information technology)	•	•	•	•			•	•		
Overall, I find the (information technology) easy to use	•	•	•	•	•	•	•		•	•

(1) Davis (1989, 1993); (2) Davis, Bagozzi and Warshaw (1989); (3) Mathieson (1991); (4) Subramanian (1994); (5) Taylor and Todd (1995a, 1995b); (6) Keil, Beranek and Konsynski (1995); (7) Szajna (1996); (8) Chau (1996); (9) Jackson, Chow and Leitch (1997); (10) Igbaria, Zinatelli et al (1997)

Perceived Ease of Use	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
I find (information technology) cumbersome to use										
Learning to operate (information technology) is easy for me		•			•	•				•
Interacting with the (information technology) is often frustrating										
I find it easy to get the (information technology) to do what I want to do	•		•	•		•	•	•	•	•
The (information technology) is rigid and inflexible to interact with			•			•	•	•		•
It is easy for me to remember how to perform tasks using the (information technology)			•							
Interacting with the (information technology) requires a lot of mental effort	•					•	•		•	•
My interaction with the (information technology) is clear and understandable		•						•	•	
I find it takes a lot of effort to become skillful at using the (information technology)						•				•
Overall, I find the (information technology) easy to use	•	•	•	•	•	•	•	•	•	•

(11) Bajaj & Nidimolu (1998); (12) Gefen & Keil (1998); (13) Agarwal & Prasad (1997, 1999); (14) Lucas & Spitler (1999); (15) Karahanna et al. (1999); (16) Hu et al (1999); (17) Dishaw & Strong (1999); (18) Venkatesh & Davis (1996, 2000); (19) Venkatesh & Morris (2000); (20) Chau & Hu (2002)

many of the reviewed investigations, a choice of student groups as a target population stands out, a quite common practice for economic reasons. This fact limits the power of generalization of their results to the population as a whole.

- As for the type of technologies that represent the social object of interchange in the examined studies, we notice how all of them center on the prediction of the use of specific technologies. The analysis of office automation computer programs or system development applications must be pointed out. The circumstance described reduces the generality of the model and questions its applicability to any situation regarding the use of technology.

- The investigations carried out propose the existence of two key factors to take into consideration when user attitude toward the use of a certain technology is assessed. However, there is a unanimous agreement among authors when they concur that it is necessary to include other components in the model so they can contribute to the explanation of the variance in the use of technology. This suggests that it would be convenient to thoroughly study the dimensionality within the technology attitude concept, redefining and improving some of the variables the various dimensions are made of, and also attempting to identify new ones.

REFERENCES

Agarwal, R., & Prasad, J. (1999). Are individual differences germane to the acceptance of new information technologies. *Decision Sciences, 30*(2), 361–391.

Ajzen, I. (1985). From intentions to actions: A theory of planned behavior. In J. Kuhl, & J. Beckmann (Eds.), *Action control: From cognition to behavior* (pp. 11–39). New York: Springer Verlag.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes. 50*, 179–221.

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Englewood Cliffs, NJ: Prentice Hall.

Ajzen, I., & Madden, T.J. (1986). Prediction of goal directed behaviours: Attitudes, intentions, and perceived behavioral control. *Journal of Experimental Social Psychology, 22*, 453–474.

Bajaj, A., & Nidimolu, S.R. (1998). A feedback model to understand information system usage. *Information and Management, 33*, 213–224.

Chau, P.Y. (1996). An empirical investigation on factors affecting the acceptance of CASE by systems developers. *Information and Management, 30*, 269–280.

Chau, P.Y., & Hu, P.J. (2002). Examining a model of information technology acceptance by individual professionals: An exploratory study. *Journal of Management Information Systems, 18*(4), 191–229.

Davis, F.D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319–340.

Davis, F.D. (1993). User acceptance of information technology: System characteristics, user perceptions, and behavioural impacts. *International Journal of Man Machine Studies, 38*, 475–487.

Davis, F.D., Bagozzi, R.P., & Warshaw, P.R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science, 35*(8), 982–1003.

Dishaw, M.T., & Strong, D.M. (1999). Extending the technology acceptance model with task-technology fit constructs. *Information and Management, 36*, 9–21.

Gattiker, U.E. (1984). Managing computer-based office information technology: A process model for management. In H. Hendrick, & O. Brown (Eds.), *Human factors in organizational design* (pp. 111–132). Amsterdam: North-Holland.

Gefen, D., & Keil, M. (1998). The impact of developer responsiveness on perceptions of usefulness and ease of use: An extension of technology acceptance model. *The DATA BASE for Advances and Information Systems, 29*(2), 35–49.

Hu, P.J., et al. (1999). Examining the technology acceptance model using physician acceptance of telemedicine technology. *Journal of Management Information Systems, 16*(2), 91–112.

Igbaria, M., et al. (1997). Personal computing acceptance factors in small firms: A structural equation model. *MIS Quarterly*, September, 279–302.

Jackson, C.M., Chow, S., & Leitch, R.A. (1997). Toward an understanding of behavioural intention to use an information system. *Decision Sciences, 28*(2), 357–389.

Karahanna, E., Straub, D.W., & Chervany, N.L. (1999). Information technology adoption across time: A cross-sectional comparison of pre-adoption and post-adoption beliefs. *MIS Quarterly, 23*(2), 183–213.

Keil, M., Beranek, P.M., & Konsynski, B.R. (1995). Usefulness and ease of use: Field study evidence regarding task considerations. *Decision Support Systems, 13*, 17–91.

Lucas, H.C. & Spitler, V.K. (1999). Technology use and performance: A field study of broker workstations. *Decision Sciences, 30*(2), 291–311.

Mathieson, K. (1991). Predicting user intentions: Comparing the technology acceptance model with the theory of planned behavior. *Information Systems Research, 2*(3), 173–191.

Morales, J.F., Rebolloso, E., & Moya, M. (1996). Modelos combinatorios del cambio de actitudes. In J.F. Morales, et al. (Eds.), *Psicología social* (pp. 555–566). Madrid: McGraw-Hill.

Muthitacharoen, A., & Palvia, P. (2002). B2C Internet commerce: A tale of two nations. *Journal of Electronic Commerce Research, 3*(4), 201–212.

Pavlou, P.A. (2003). Consumer acceptance of electronic commerce: Integrating trust and risk with the technology acceptance model. *International Journal of Electronic Commerce, 7*(3), 101–134.

Subramanian, G.H. (1994). A replication of perceived usefulness and perceived ease of use measurement. *Decision Sciences, 25*(5/6), 863–874.

Szajna, B. (1996). Empirical evaluation of the revised technology acceptance model. *Management Science, 42*(1), 85–92.

Taylor, S., & Todd, P.A. (1995a). Understanding information technology usage: A test of competing models. *Information Systems Research, 6*(2), 144–176.

Taylor, S., & Todd, P.A. (1995b). Assessing IT usage: The role of prior experience. *MIS Quarterly*, December, 561–570.

Venkatesh, V., & Davis, F.D. (1996). A model of antecedents of perceived ease of use: Development and test. *Decision Sciences, 27*(3), 451–481.

Venkatesh, V., & Davis, F.D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science, 46*(2), 186–204.

Venkatesh, V., & Morris, M.G. (2000). Why do not men ever stop to ask for directions? Gender, social influence, and their role in technology acceptance and usage behaviour. *MIS Quarterly, 24*(1), 115–139.

KEY TERMS

Acceptance of Technology: Psychological state of individuals with regard to their voluntary or desired use of a particular technology**.**

Attitude: A state of mind or a feeling.

Ease of Use: Refers to the property of a product or thing that a user can operate without having to overcome a steep learning curve.

Expectancy-Value Behavior: Existence of conducts that, albeit in part, a person cannot voluntarily keep under control.

Perceived Behavioral Control: Refers to people's perceptions of their ability to perform a given behavior.

Technology Readiness: Refers to a combination of beliefs related to technology that collectively determine the inclination in a customer, employee, or executive to adopt new technologies in order to reach his or her objectives, both at work and during leisure time.

Usefulness: The quality of being suitable or adaptable to an end.

Telemedicine and Information Technology for Disaster Medical Scenarios

John D. Haynes
Charles Darwin University, Australia

Mehnaz Saleem
Sir Syed University of Engineering and Technology, Pakistan

Moona Kanwal
Sir Syed University of Engineering and Technology, Pakistan

INTRODUCTION

Disasters constitute events which are catastrophic in nature. Such events critically threaten the health, safety, and lives of people and their environment (and even aspects of the global environment), and as a result, overwhelm the affected community's emergency response capacity. Globally, a major disaster occurs almost daily. Consequently, disaster events are virtually an everyday fact of life. Emergency medical services constitute one important aspect of disaster responses. Those populations affected by disasters require a complete range of health services and the appropriate mechanism of delivery. In this respect, increasingly, information technology is playing a greater role. Disaster medicine has become more than merely a mass-casualty, and affected health response; the affected population's needs are assessed, which range from medical requirements, to rapidly coordinating and providing casualty, routine, and preventive health services. These kinds of assessments are significantly more effective, given the appropriate deployment of current information technology.

There are three major time phases associated with disaster response (Patoli, 2005). In the Predisaster Phase, emphasis is placed on activities which address prevention and preparedness, which entail hazard and vulnerability assessments, human and material resource inventory, comprehensive planning, and exercises to test disaster and recovery plans, capabilities, and skills.

In the Disaster Relief Operations Phase (hours to weeks), assessment of damage and consequent recovery requirements, and the planning and implementation of emergency plans, have priority.

The Postdisaster Rehabilitation Phase may extend for months, or even years, as management, maintenance, and infrastructure, and various community activities are rehabilitated and restored.

Telemedicine can be integrated into all of the actives of medical disaster management. This article will explore the role of telemedicine in (medical) disaster management. The use of communications pathways—information technologies which include global positioning system (GPS) technologies, hypermedia, artificial intelligence, software and hardware for computer miniaturization, and advanced wireless sensor systems—in telemedicine will be discussed with respect to their advantages, disadvantages, and future potential.

WHAT IS TELEMEDICINE?

Scott (1996) indicates that telemedicine is "just a new label for applications of telecommunications and information technologies [applied to the area of medicine]." But more specifically, telemedicine can be defined as diagnosis, consultation, treatment, education, and the transfer of medical data (i.e., high resolution images, sounds, live video, and patient records) using interactive audiovisual and data communications (Jerant, Schlachta, Epperly, & Camp, 1998). What this means is that a physician located far from a reference center can consult his colleagues remotely in order to solve a difficult case, supplemented with continuing education course over the Internet, or by accessing medical information from digital libraries.

From Sunday and Pothier (2001), "telemedicine isn't a new medical specialty; it simply offers another way to deliver existing services to people who can benefit from them no matter where they are." Sunday and Pothier note that in the Telemedicine Report to U.S. Congress, 2001, telemedicine was defined as "the use of electronic communication and information technologies to provide or support clinical cases at a distance" (Sunday & Pothier, 2001). These same tools can also be used to facilitate exchanges between centers of medical expertise, at a national or international level (Geissbuhler, Ly, Lovis, & L'Haire, 2003).

In terms of its commonly understood sense, in which a doctorpatient interaction involves telecommunication, historically, this goes back at least to the use of ship-to-shore radio for giving medical advice to sea captains. A few years ago, the term "telemedicine" began to be supplanted by the term "telehealth," which was thought to be more "politically correct," but recently, this too has been overtaken by even more fashionable terms, such as "online health" and "ehealth." (Wootton, 2001)

Telemedicine in Medical Disaster Management

As noted in brief in the Introduction, Patoli tells us (Patoli, 2005) that telemedicine in medical disaster management is divided into three areas. We will now explore those areas in more detail:

- **Phase 1: Predisaster Preparedness.** The first consideration is planning and development, and technological needs. In this planning stage, Patoli (2005) indicates there are six categories of information that are crucial: management, clinical, surveillance and epidemiological, literature in terms of documentation and reports, diagnostic knowledge, personal, and community knowledge. The second consideration is the human factor. Patoli (2005) indicates that the human factor is broken into four interrelated areas: management and administration of the telemedicine system, medical expertise and other health and medical activities, patients and telemedicine users, and e-training. Four more major areas for Phase 1 follow: databases, artificial intelligence, expert systems and decision support systems, geographical information systems, and research projects (Patoli, 2005).
- **Phase 2: Disaster Relief Operations.** In this phase, responses are activated as immediately as possible after the disaster is detected. The main areas in this phase are: robot assisted medical reachback, patient tracking systems, critical state patient's telemonitoring through sensor devices, prehospital management through mobile technology, telediagnostics entailing both teleradiology and telepathology, clinical decision support systems, and finally, telesurgery and teleconsultation (Patoli, 2005).
- **Phase 3: Postdisaster Rehabilitation.** In this phase, many of the disaster victims live in camps and temporary shelters, suffering from depression and stress. The main areas in this phase are: telepsychiatric intervention, public health threats entailing early warning systems, disaster medicine, epidemiology, and telerehabilitation (Patoli, 2005).

APPLICATIONS OF TELEMEDICINE

To summarize from Tracy (2004):

Telehealth and telemedicine are becoming more and more engrained in the delivery of everyday healthcare, distance education and health care administration. Extensive numbers of patients in underserved areas are receiving services they may not have otherwise received without traveling great distances or overcoming other transportation barriers. Telemedicine services (from the youngest of patients to the frail elderly) range from primary care, to the highly specialized care found in leading academic medical centers. Telehealth systems can be found in hospitals, clinics, nursing homes, rehabilitation hospitals, homes, assisted living facilities, schools, prisons, or health departments. In fact, today you may find telehealth systems any place healthcare is provided. (Tracy, 2004)

TELEMEDICINE IN DISASTERS AND EMERGENCIES

Garshnek and Burkle (2000) very articulately sum up:

Emergency and Disaster events are the most important scenarios where telemedicine can best be utilized. Success or failure of a disaster response often is determined by timely access to communication and reliable information. The rapid progress and future course in telecommunications indicate that lack of communications no longer needs be the paralyzing factor in a disaster scenario. This is important especially for medical response where time is of essence to save lives.

What is most urgent in the events of medical disasters and emergencies is the need for smart deployment of systems and procedures that provide for the accurate assembly of personnel, equipment, and telecommunications resources. The recent hurricanes in USA alone, let alone recent tsunami and earthquake disasters, have acutely sharpened awareness in this respect. More specifically, in a recent statement from the USA Homeland Security (http://www.gtp-etc.com/docs/ECU%20Homeland%20Security.doc):

These [types of] experience have stimulated the Telemedicine Center to address the development of capabilities that include: (1) rapidly deployable and reconfigurable medical device and sensor toolkits; (2) on demand, geography-independent telecommunications resources; and (3) tools to improve field responders' capabilities in emergency or disaster situations.

Indeed, disaster telemedicine systems do not need to be special or sophisticated. We again reiterate that the challenge is to match the right systems with a appropriate disaster plan or scenario (Garshnek & Burkle, 1999). In almost all cases with natural disasters which are wide-ranging, the basic infrastructure of the region is dislocated and largely unusable. Therefore, without ground communications, special considerations need to be enacted.

In the past, providing telemedicine capabilities to local disaster areas has been expensive. To establish land line or satellite linkages, often bulky equipment needs to be deployed. Traditionally, only large governments and commercial enterprises have been able to acquire such systems and infrastructures. (Garshnek & Burkle, 1999)

Consequently, "satellites do play a significant role for mobility and land-line independence for telemedicine" (Garshnek & Burkle, 1999). Flexibility such as this is especially important. However, as indicated above, the advantages of satellites are often outweighed by their disadvantage in terms of initial costs and maintenance costs. The good news is that portable satellite terminals, now being evolved, will be less expensive in the very near future (Garshnek & Burkle, 1999).

Garshnek and Burkle (1999) succinctly note the following:

Real-time video teleconferencing of a sophisticated nature was not always required or useful in disaster [scenarios]. The time flexibility of store-and-forward teleconsultation has now been appreciated and if often necessary where real-time cases are not needed. Internet and multimedia e-mail are evolving. In complicated disaster scenarios, easy-to-use [and portable] Internet-based consultation systems might well be ideal.

TELEMEDICINE AND INTERNATIONAL DISASTER RESPONSE

Furthermore, Garshnek and Burkle (1999) note the following:

While many information management and telecommunications technologies are currently employed in disaster response, there are few reports of telemedicine being utilized in disaster settings. The Pan American Health Organization (PAHO) has provided satellite communication ground stations to support disaster response. The longest and most extensive use of telemedicine was the NASA-Russia Space Bridge employed during the Post-Disaster Rehabilitation Phase after the devastating Armenian Earthquake of l988.

In relation to the Space Bridge Project, which used satellite communications to provide medical consultation to several Armenian regional hospitals, Scott (1996) noted that:

in a separate link, consultation was also provided to the Russian town of Ufa, where a gas explosion during this same period of time caused a large number of casualties. Slow-scan black and white video was transmitted from Ufa to one of the Space Bridge sites in Armenia (Yerevan) which provided satellite uplink.

Consequently, Telemedicine was tested and deployed for the first time on a large scale in the Space Bridge Project, in such a way that the technology utilized was proven very effective.

TELEMEDICINE COMMUNICATIONS PATHWAYS

As indicated above, from Garshnek and Burkle (1999), a correct match needs to be effected between the medical disaster plan or scenario, and the communication and contingency system. In this respect, wireless transmissions, which utilize the entire frequency spectrum, are advantageous, especially when it is considered that disaster damage has little affect on wireless transmissions. However, when there is damage to repeater stations, satellites offer better long distance communications than other means. And portable satellites offer great advantages.

INFORMATION TECHNOLOGIES FOR TELEMEDICINE

There are a whole range of information technologies available for telemedicine. This article will focus upon GPS technologies, hypermedia, artificial intelligence (in particular artificial neural networks), software and hardware for computer miniaturization, and advanced wireless sensor networks.

GPS Technologies for Telemedicine

In relation to GPS technologies, Finnegan (School of Information Technology, Charles Darwin University) is currently working on extending GPS technology into everyday use of mobile computer technologies, particularly supporting project and other work in remote locations. The intent is to provide context specific information, time, and location, which would be significantly important for telemedicine. This development will be leveraged by the emerging practice of embedding GPS into (portable) wireless laptops (Finegan, 2006).

Hypermedia for Telemedicine

Another important information technology for telemedicine is Hypermedia Documentation. Hypermedia can briefly be described in terms of the presence of information types such as multimedia (text, image, video, and audio), and the mapping of semantic connectors or links within these various multimedia types. In a paper by Saleem and Kanwal, they state the following: "hypermedia documentation schemes can be efficiently used in telemedicine development and implementation, especially in the case of scarce resources with an additional advantage of economy as compared to primitive databases" (Saleem & Kanwal, 2006).

Artificial Intelligence for Telemedicine

Much has been written in the literature concerning the relation between artificial intelligence and telemedicine, but the literature is relatively silent on artificial neural networks. Artificial neural networks (ANNs) offer potentially great advantages, among other considerations, such as pattern recognition generally, forecasting, and optimisation, not to mention the provision for early warning disaster scenario detection.

As Kumar and Haynes (2004, pp. 7) note:

ANN systems can be used to assist actively in a number of problems faced in the operations research area and statistical like: Prediction/ estimation (curve fittings, forecasting etc); pattern recognition (classification, discrimination analysis and diagnosis); clustering (grouping without a prior knowledge of classes); optimization (the solution of linear and non linear mathematical programming model involving continuous and/or discrete variables). Pattern recognition related decision problems suit artificial neural networks more than to statistical techniques primarily because of the complex nature of problems, which are not well understood mathematically and involve subjectivity. In addition, such problems have qualitative and noise data. Even if the values of every input feature are not known, a trained neural network will produce a response.

Consequently, artificial neural networks are very robust—that is, their input data can vary and the ANN will still be able to process, or train upon the data, to make a "conclusion," or more precisely, an output. And in the view of the authors, they would be ideal in a medical disaster situation where precise inputs are not readily available.

One current disadvantage of an ANN, even with portable PCs, is the time it takes to collect data; but, with advances in online real-time collection techniques, the advantages and future potential of ANNs is very promising for telemedicine.

Telemedicine and Software and Hardware for Computer Miniaturization

Perhaps the most interesting example of computer miniaturisation, and one of the most recent for telemedicine, is the scalable software infrastructure for miniature medical devises, known as CodeBlue (CodeBlue, 2006). The design of CodeBlue provides for "routing, naming, discovery, and security for wireless sensors, PDAs, PCs, and other devices" (CodeBlue, 2006), for the treatment and monitoring of medical patients.

In terms of the hardware computer miniaturisation, CodeBlue utilises SHIMMER, which incorporates a TI MSP430 processor, a radio, a triaxial accelerator, and a rechargeable battery, all in miniature. The latest SHIMMER is three times the size of an Australian twenty-cent piece. Further,

Shimmer includes a MicroSD slot supporting up to 2 GBytes of flash memory. This allows SHIMMER to store significant amounts of data (2GB can store up to 80 days of continuous triaxial accelerometer datas sampled at 50Hz. Shimmer can also be configured with an optional Bluetooth radio. (CodeBlue, 2006)

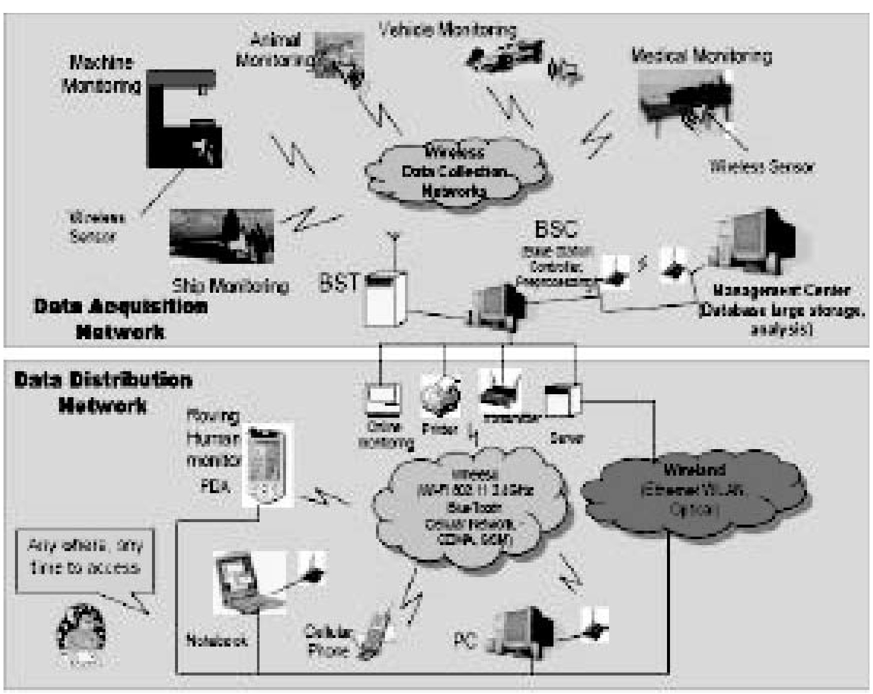
Advanced Wireless Sensor Systems for Telemedicine

In 1980, modern research commenced at DARPA in the Distributed Sensor Networks Program (DSN) (Bobbie, Deosthale, & Thain, 2006). Today, small and inexpensive MEMS technology sensors, wireless technology and low-power processors provide excellent support for telemedicine. A typical wireless sensor network can be seen in Figure 1 (Lewis, 2004).

For disaster medical scenarios, the advantages for wireless sensor networks are relative inexpensiveness, portability, and nonreliance on land-lines. However, a disadvantage can be where wireless repeater stations are damaged by the disaster, in which case satellites are preferred.

It is not the aim of this article to go into a detailed analysis of wireless sensor networks. More appropriate

Figure 1. A typical wireless sensor network (Lewis, 2004)



are a series of recent applications of wireless sensor networks for telemedicine. Bobbie et al. (2006) discuss four such applications:

1. **Blue Nurse:** This was developed by the University of Queensland; its design entails point-to-point communication to a single bluetooth module for the purpose of receiving ECG date on a PC. As mentioned by Bobbie et al. (2006), the reference is Daniel Marr, ECG Application Featuring Data transmission by Bluetooth Ph.D. thesis, University of Queensland, October, 2001.
2. **Proactive Health:** This is a system for the daily activities of the elderly entailing postage sized wireless radio frequency identification tags (RFID), where data is communicated back to a central system. Bobbie et al., (2006) indicate the reference as: http://www.cms.hhs.gov/hipaa.
3. **Blue Medica:** This entails a tablet PC viewing data from a hospital information system by a doctor, where a wrist band carrying the sensor was used to measure conditions, such as temperature. Bobbie et al. (2006) give the references as Kostelnik, R., and Lacko, P. Blue medica – wireless medical data access appliance. *Journal of Electrical Engineering, 52*(9–10), 311–315.
4. **NSF COPIRE:** Provides a system for cardiology evolution, where a sensor communicates data with a PC.

Stankovic et al. (2005) discuss wireless sensor networks in terms of their potentials and their challenges. They state that wireless sensor networks for the future must be able to address each of the following medical applications: integration within existing medical practices and technology, real-time, long-term, remote monitoring, miniature wearable (and durable) sensors, and assistance to the elderly and chronic patients (Stankovik et al., 2005). From our perspective in terms of medical disaster scenarios, clearly, victims of disasters, in a sense, fall into the category of the embodiment of "degrees of helplessness" (such as homelessness, depression, and acute stress) to the elderly and the chronic class of patients. They state that the Wireless Sensor Networks can easily display their use in terms of (1) sleep apnea (2) journaling support, particularly rheumatic diseases—but, for our purposes in medical disaster scenarios, the patient under journaling support records changes in bodily functions (pain, sleep, headache, irritability, and so on), (3) cardiac health, especially cardiac arrhythmia and cardiac stress.

The (USA) National Science Foundation (NSF) indicate that the future potential of sensors and sensor systems need to include effective responses to (1) toxic chemicals, explosives, and biological agents, (2) enhanced sensitivity and fewer false alarms, and (3) unattended functionality in complex environments (NSF, http://www.nsf.gov/news/special_reports/sensor/nsf_sensor.jsp). The NSF, to indicate a few they mention, suggest that genomic tools may be incorporated to aid in detecting ecological changes with the support of nano- and meso-scale complex miniature hardware structures; arrays of ultra low-power wireless nodes need to be incorporated into high bandwidth reconfigurable networks (refer to NSF Web site).

THE INTEGRATION OF INFORMATION TECHNOLOGY AND TELEMEDICINE

It is suggested that once a medical disaster management has been implemented in a country or a site that it become a learning event for that community. If it can happen once, it can happen again. Therefore, preparedness is the key. Preparedness entails learning and preparing for the future on a larger scale entails community involvement, hence the key becomes a community learning culture focused towards the possibility of another medical disaster. Haynes (Haynes, 2001, pp. 17) notes in this respect:

the learning culture must evolve [for itself], if it were to remain static or stagnate it would die as a culture. A learning culture is alive in the sense that it is intimately connected with the humans that draw strength from it and who, by having a deep sense of it, keep it alive.

The further key, then, is to instill in a community a sense of what it entails to be a learning culture; in this case, a learning culture directed towards medical and community disaster management and what information technologies to effectively implement, based on their continued use and development and to provide effective plans for these information technologies integration into the societies concerned.

A crucial point made by Saleem and Kanwal (2007) is that the development and evolution of telemedicine should be "need driven and not technology driven."

Clearly, many disadvantages can be enumerated in respect of telemedicine for underdeveloped countries, but with appropriate polices and the realization that telemedicine systems must be viewed by practitioners as useful to them, the advantages and an exciting future potential for telemedicine and information technology can be realized.

CONCLUSION

Disasters are virtually, globally, a daily occurrence. Telemedicine can be utilized in disasters only within the constraints of local and regional infrastructure capabilities. But three phases are common to all medical disaster events: predisaster; disaster relief operations; and postdisaster rehabilitation. We have considered what telemedicine is defined as, and perhaps the most succinct definition is the utilization of "electronic communication and information technologies to provide or support clinical cases at a distance" from the Telemedicine Report to U.S. Congress, 2001. We have explored issues associated with the role of telemedicine in disaster management in terms of the use of communications pathways—information technologies which have included GPS systems, hypermedia, artificial intelligence, software and hardware for computer miniaturization, and advanced wireless sensor systems. We have examined their respective advantages, disadvantages, and future potential. Perhaps the most promising of these information technologies are artificial neural networks and the development of advanced wireless sensor networks in conjunction with computer miniaturization.

Two principal recommendations from this article are: first, that if a disaster can happen in one place once, then it can happen in the same place again; therefore, preparedness needs to be paramount. This entails the evolvement of a learning culture focused upon what to do for future medical disaster events and what appropriate information technologies need to be implemented. These information technologies must be developed in terms of their continued use with the necessary provision for effective plans for their continued evolvement. Second, that telemedicine must be need-driven and not technology driven. It follows that community needs should be at the foundation of the development of information technologies, not only in relation to medical disaster management, but also for humanity as a species.

REFERENCES

Bobbie, P.O., Deosthale, C., & Thain, W. (2006, March–May). Telemedicine: A mote-based data acquisition system for real time health monitoring. *Telehealth*. Retrieved February 5, 2008, from http://www.actapress.com/Content_Of_Proceeding.aspx?ProceedingID=386

CodeBlue. (2006). *Wireless sensor networks for medical care*. Division of Engineering and Medical Sciences, Harvard University. Retrieved February 5, 2008, from http://www.eecs.hardvard.edu/~mdw/proj/codeblue/

Finegan, A. (2006). *Clever technologies – filling the gap between intent and realization.* Infrastructure Grant, School of Information Technology, Charles Darwin University, Australia. Retrieved February 5, 2008, from http://www.cdu.edu.au/it/senior-research.html

Garshnek, V., & Burkle, F.M. (1999, January–February). Applications of telemedicine and telecommunications to disaster medicine: Historical and future perspectives. *JAMIA Medical Information Association, 6*(1), 26–37. Retrieved February 5, 2008, from http://www.jamia.org/cgi/reprint/6/1/26.pdf

Garshnek, V., & Burkle, F.M. (2000). *Disaster telemedicine - part II: Future applications of telemedicine and telecommunications in disaster medicine*. Retrieved February 5, 2008, from http://pdm.medicine.wisc.edu/garshnek.htm

Geissbuhler, A., Ly, O., Lovis, C., & L'Haire, J. -F. (2003, November). Telemedicine in Western Africa: Lessons learned from a pilot project in Mali, perspectives and recommendations. In *Proceedings of AMIA Annual Symposium* (Vol. 2003**,** pp. 249–253). Retrieved February 5, 2008, from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1479936

Haynes, J.D. (2001). Churchman's hegelian inquiring system and perspectival thinking. *Journal of Information Systems Frontiers, 3*(1), 29–39. Netherlands: Kluwer Academic Publishers.

Jerant, A.F., Schlachta, L., Epperly, T.D., & Camp, J.B. (1998, January). Back to the future: The telemedicine house call. *Family Practice Management, 5*(1). Retrieved February 5, 2008, from http://www.aafp.org/fpm/980100fm/lead.html

Kumar, K., & Haynes, J.D. (2004, May). Forecasting credit ratings using an ANN and statistical techniques. *International Journal of Business Studies,* May.

Lewis, F.L. (2004). *Technologies, protocols and applications: Smart environments.* In D.J. Cook, & S.K. Das. New York: John Wiley.

Patoli, A.Q. (2005, December). Role of telemedicine in disaster management. *eHealth International Journal, 2*(2). Retrieved February 5, 2008, from http://www.ehealthinternational.net/

Saleem, M., & Kanwal, M. (2006, March). Hypermedia document structure to support telemedicine in local environment. *International Conference on Software Engineering (ISE'06).* Lahore, Pakistan.

Saleem, M., & Kanwal, M. (2007, January). Incorporating telemedicine into local environment: Uses and challenges. *5th International Bhurban Conference on Applied Science and Technology (IBCAST-2007).* Islamabad, Pakistan.

Scott, J.C. (1996, June). Telemedicine in disaster applications. *UN/ESA/Chile Workshop on Space Technology to Prevent and Mitigate the Effects of Disasters for the Benefit of Developing Countries.* Retrieved February 5, 2008, from http://www.uku.fi/english/organizations/helina-l/msglog/1996/0019.html

Stankovic, J.A., et al. (2005, June 2–3). Wireless sensor networks for in-home healthcare: Potential and challenges. In *High Confidence Medical Device Software and Systems (HCMDSS) workshop.* Philadelphia, PA.

Sunday, J., & Pothier, P. (2001). AT & T Inc. Retrieved February 5, 2008, from http://www.corp.att.com/healthcare/docs/article_telemedicine.pdf

Tracy, J. (2004). *Telemedicine technical assistance documents - a guide to getting started in telemedicine*. Forward by Dean Puskin, Office for the Advancement of Telehealth (pp. 400). University of Missouri, School of Medicine.

Wootton, R. (2001, September). Clinical review: Recent advances telemedicine. *BMJ*, 323, 557–560.

KEY TERMS

Artificial Neural Networks: Computer systems (hardware and software) that "learn" by training on inputs, largely with pattern recognition as their basis of operation.

Hypermedia: Information types such as multimedia (combination and integration of text, image, video, and audio).

Need-Driven Telemedicine: Information technology that is developed with the benefit of the community in mind, in relation to telemedicine.

Medical Disaster Management: A disaster plan for medical disasters consisting of three phases: predisaster, disaster relief operations, and postdisaster rehabilitation.

Telehealth: Synonymous with telemedicine.

Telemedicine: Electronic communication and information technologies to provide or support clinical cases at a distance.

A Theoretical Model of Observed Health Benefits of PACS Implementation

Lucy Firth
University of Melbourne, Australia

Reeva Lederman
University of Melbourne, Australia

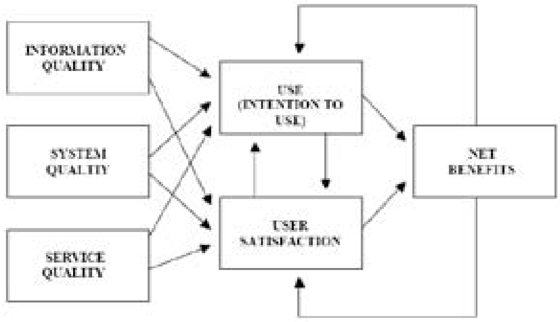
Mazen Ali
University of Melbourne, Australia

INTRODUCTION

Picture Archiving and Communication Systems (PACS) is an important e-health application, playing a significant operational role within hospitals in electronically transmitting image-based data. Various authors have attributed a range of benefits to PACS, including diagnostic accuracy (Scott et al., 1995; Slasky et al., 1990), interpretation time savings (Kato et al., 1995), workflow (Gale, Gale, Schwartz, Muse & Walker, 2000) and user satisfaction (Philling, 1999). While there is general consensus that PACS brings some or all of these benefits, there is little agreement as to the mechanism through which these benefits are generated.

The work of Delone and McLean (2003) provides a generalized model that attributes information systems success to benefits accruing from relationships between characteristics of the information system (e.g., information quality, system quality, service quality) and user intention to use and user satisfaction (see Figure 1).

Figure 1. Delone and McLean's (2003) model of IS success



This chapter provides empirical evidence that the Delone and McLean (2003) model with modifications explains the source of PACS benefits. Not only do we suggest a new model, but we also recommend that each specific information system context may require specific modification of Delone and McLean's (2003) model to explain the source of success.

BACKGROUND

This model and its earlier incarnation (Delone & McLean, 1992) have been applied to a range of IS contexts, including PACS (Cox & Dawe, 2002; Reiner et al., 2002). While this indicates that Delone and Mclean's (2003) model is robust when applied to PACS, it appears from the literature that other factors are causally related to the net benefits attributed to PACS. In particular, the literature suggests the following:

- The environment in which images are viewed and interpreted (Abdullah & Ng, 2001; Rostenberg, 2006; Reiner & Seigel, 2000) impacts the speed, accuracy, and comfort of interpretation. Rumreich and Johnson (2003) specifically argue that a purpose-designed reading environment improves radiologist satisfaction and efficiency.
- Training offered to users is also said to be important to the benefits from PACS (Protopapas et al., 1996; Watkins, Weatherburn & Bryan, 2000; Law & Zhou, 2003; Huang, 2003). Not only does training improve performance (Barney, 1991; Ulrich, 1991), but the way it is conducted also

increases user satisfaction (Bradford & Florian, 2003).

This suggests that while Delone and McLean's (2003) model may be valuably applied to PACS environments without modification, the inclusion of environmental and training issues may better explain the source of benefits and, thus, IS success.

IDENTIFYING AND EXPLAINING SOURCES OF PACS SUCCESS

The identification of the nature and source of benefits from PACS involved a single-stage case study of 18 PACS users, administrators, and managers at a large private hospital. The hospital has more than 60 departments and service centers, 500 beds, and nine operating theaters. The Medical Imaging Department is equipped with all imaging modalities and provides radiological services throughout the hospital. A mini-PACS was implemented in 1995 covering the intensive care, emergency, and imaging departments. In 2000, a decision was made to implement PACS hospitalwide, and in 2004, this was achieved.

This investigation asked the following: Are the two extra elements (image interpreting environment training) considered to be of importance at the hospital, and would their inclusion improve the relevance of Delone and McLean's 2003 model to the PACS environment? The investigation was part of a larger project that sought to evaluate the recent PACS implementation from three benefit perspectives: the impact on workflow; the impact on timeliness of image delivery and reporting; and the impact on patient throughput. The work presented here cuts across each of these perspectives, to focus on the aim of understanding the nature and source of benefits of PACS.

The investigation had both qualitative and quantitative components. Semistructured interviews around the topics of benefits (perceived or formally evaluated), system quality, service quality, training, viewing environment issues, and the attainment of hospital goals. Each interview was followed by a quantitative survey that was administered by the researcher. The survey addressed issues of self-efficacy and intention to use PACS, training, support, and level of user satisfaction. These questions were drawn from scales developed by Speier and Venkatesh (2002). The interview, survey, and documents were analyzed using open coding to divide the data into concepts and categories of concepts, and to assign properties to those categories (Miles & Huberman, 1994). Following this, axial coding was used to organize the categories and dimensions into groupings of ideas that are thematically related (Strauss, 1987). The quantitative survey data were analyzed descriptively only as the sample was too small for more sophisticated analyses.

The major findings and the modifications that they suggest for Delone and McLean's (2003) model follow:

- Adoption of PACS was universal with no interviewee indicating that he or she still used or preferred film-based images where there is a PACS digital browser-based alternative.
- Users universally reported not only that they were satisfied with PACS, but also that the more they used PACS, the more satisfied they became and the greater intention they had to use PACS' functionality. That is, use/intention to use and satisfaction with PACS are recursively promoted by the benefits that the users perceive. The model would be improved by arrows indicating a reverse loop from benefits to use and satisfaction (see Figure 2).
- DeLone and McLean's (2003) suggestion that where use (intention to use) and user satisfaction exist, net benefits will follow is confirmed. No modification of the model is needed.
- "Information quality" could be meaningfully relabeled "information and image quality" to reflect the multimedia nature of PACS presentation of data.
- For some respondents, information and image quality is directly related to benefits as well as indirectly via user perceptions and responses as Delone and McLean (2003) suggest. This direct benefit was not universally reported. Those that did not report this direct benefit commented that image quality was better, notwithstanding the trade-off between contrast and resolution, but that this benefit was entirely accounted for by the indirect effect through intention to use. Those that did report a direct effect included orthopedic surgeons who commented that because the size of the digital image is not necessarily related to

the object of interest, they experience direct disbenefits, as indicated by the following quote:

In orthopaedics surgery we focus on joint replacements, we like to be able to look at an X-ray and say we know how big that patient's knee is and how big his hips and that's where one of the biggest draw backs is with the computerized X-ray system. ... So there is an inefficiency in terms of planning for some operations and not able to easily put up plastic templates or plastic picture or a transparent picture of a joint replacement. Put that over an X-ray and say that's the correct size… Now if I put a ruler on this on computer that will measure anything between 29 and 32 ml and that leaves you with 2-3 ml.

This finding suggests that the model be modified to incorporate the possibility of a forward loop direct from information and image quality to benefits (see Figure 2).

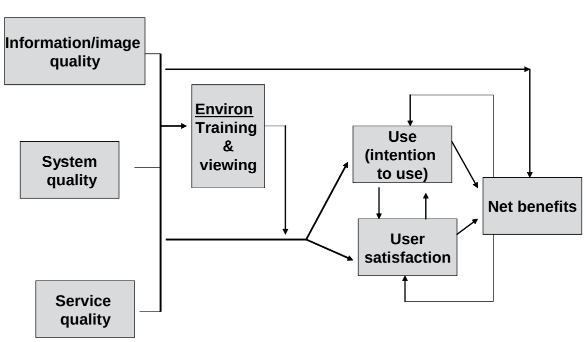
- While DeLone and Mclean (2003) found a direct relationship between the system characteristics and user perceptions and responses, our finding suggests that this relationship is both direct and indirect. The relationships are indirect to the extent that they are intermediated by two environmental factors that play an active role not normally attributed to environmental factors. The two environmental intermediators are training and the image viewing environment. In our findings, the training that impacted users' levels of satisfaction and intention to use was not training received by that user but by those in the environment upon whom that user relied in order to perform his or her own tasks. Our data show that user satisfaction and users' sense of their own productivity in using PACS were enhanced by the training that had been received by others. Interviewees reported that, where their colleagues had been inadequately trained, the effect on their own user satisfaction and on benefits in terms of productivity were strongly negative. The data also show that a lack of training of colleagues made it more difficult for users to gain the benefit of the information and image quality, and of the system quality. The data do not indicate that the user's ability to gain the benefit from service quality was impaired by environmental factors. While no interviewee reported intentions not to use the system because of the impact of the training and consequent competence (or lack thereof) of others, the impacts on user satisfaction and on benefits are captured in this quote by a surgeon:

There are a lot of people that don't know how to use it all, nursing staff etc. so we are operating a scrub or unable to register a computer screen. In the past you just say "can you put up that Xray," for the nursing staff who were not educated in using the system it's a big deal. It's very difficult for them, and it's very difficult to explain as well without touching anything.

This finding that training of colleagues plays on important role in intermediating between the characteristics of the system and the user satisfaction and benefits attributed to that system suggests that training of colleagues be introduced to the model as a possible mediator between system characteristics and user attributes (see Figure 2).

- If the benefits of image manipulation and contrast are to be enjoyed, the image interpreting environment has to be suitable in terms of lighting, screen size, and monitor quality. Moreover, the monitors need to be located appropriately, given the ward layout and workstation to ensure that images can be quickly and easily accessed. The fact that several clinicians might share one monitor was not considered to be a problem. This is partly because they have been accustomed to sharing light boards. The importance of the image interpreting

Figure 2. Model of PACS success



environment to both the user's intention to use and his or her satisfaction was especially apparent to those who were familiar with a previous mini-PACS implementation in only the emergency and intensive care wards. The mini-PACS had the viewing monitors isolated from the ward in the old reading room. While the number of walking steps from the current location of the monitors in the ward to the old viewing room location was as few as 10, the clinicians reported that it had been problematic. The implicit disruptions to workflow coupled with the poor viewing conditions not only led to inefficiencies, but also reduced the level of satisfaction and intention to use. This suggests that the model be modified to include the viewing environment as a possible mediator between system characteristics and user attributes (see Figure 2).

These findings suggest the model for information systems success in the PACS environment depicted in Figure 2. The model presents an extension to, rather than a departure from, Delone and McLean (2003). The model should be read as saying that as well as the expected relationships among the three characteristics of the system and the user response, there are intermediating environmental factors (i.e., training of others and viewing) playing an active role. Also, as well as the indirect relationship between the system characteristics and the net benefits, there are direct net benefits that are not mediated by intention to use and by user satisfaction.

FUTURE TRENDS

The need to evaluate health care information systems indicates a need for models that suit the health care environment specifically. Because of the differences in health contexts and conventional business contexts, this may mean customizing models validated in other contexts. This may be the case particularly in evaluating health care information systems used in not-for-profit, a service to underserved communities (including remote, impoverished, and elderly). The work presented here is evidence that it is possible to develop models that capture the uniqueness of health contexts while piggybacking on the work done for conventional business contexts. The growth of interest in informatics to provide relief in least developed nations (especially African) indicates a future trend toward model modification. Such modification is to be recommended if appropriate, relevant, and powerful evaluations are to be undertaken.

CONCLUSION

The first conclusion is that although Delone and McLean's (2003) model has the flexibility and robustness to apply to PACS contexts, its relevance may be enhanced by the inclusion of context-specific factors. In the case of explaining the source of benefits from PACS, two factors that should be included are training of others and the image interpreting environment.

Second, the inclusion of mediating conditions, both environmental, make Delone and McLean's (2003) model less primitive, while not destroying its simplicity. It would be valuable to have a quantitative measure of the strengths of the relationships as offered by a path analysis. The present study did not collect data suited to such an analysis, which is at once a limitation of this study and a proposed future direction.

The third conclusion is that training of colleagues plays a direct role in intermediating between system characteristics and use/intention to use. This direct role for environmental factors may be common to contexts in which knowledge workers rely on each other and on the information system in stressful conditions. The orthopedic theater is a particularly good example.

Fourth, we need to constantly be aware that the technology may matter. In this example, we see that the technology (PACS) is of central importance to the design of the viewing environment. Locating, lighting, and accessing new technologies in the same way that we did previous technologies may reduce the net benefits of the new technologies. In the extreme, it could make for net disbenefits.

Therefore, the model of PACS success, as developed in this project, shows how a theoretical approach can add real insight to practice. It provides an academic explanation of the practical question regarding the source of observed and measured benefits of PACS implementations.

REFERENCES

Abdullah, B.J., & Ng, K.H. (2001). In the eyes of the beholder: What we see is not what we get. *British Journal of Radiology, 74*, 675–676.

Agarwal, R., & Karahanna, E. (2000). Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage. *MIS Quarterly, 24*(4), 665–694.

Barney, J. (1991). Firm resources and sustained competitive advantage. *Journal of Management, 17*, 99–120.

Bradford, M., & Florian, J. (2003). Examining the role of innovation diffusion factors on the implementation success of enterprise resource planning systems. *International Journal of Accounting Information Systems, 4*, 205–213.

Cox, B. & Dawe, N. (2002). Evaluation of the impact of a PACS system on an intensive care unit. *Journal of Management in Medicine*, *16*(2-3), 199-205.

Delone, W.H., & McLean, E.R. (1992). Information systems success: The quest for the independent variable. *Information System Research, 3*, 60–95.

Delone, W.H., & McLean, E.R. (2003). The DeLone and McLean model of information systems success: A ten year update. *Journal of Management Information Systems, 19*, 9–30.

Gale, D., Gale, M., Schwartz, R., Muse, V., & Walker, R. (2000). An automated PACS workstation interface. *American Journal of Roentgenology, 174*, 33–36.

Huang, C. (2003). Changing learning environments with new interactive and media rich instruction environments: Virtual lab case. *Computerized Medical Imaging and Graphics, 27*, 157–253.

Kato, H., et al. (1995). Preliminary time-flow study: Comparison of interpretation times between PACS workstations and films. *Computerized Medical Imaging and Graphics, 19*, 261–265.

Law, M.Y., & Zhou, Z. (2003). New direction in PACS education and training. *Computerized Medical Imaging and Graphics, 27*, 147–156.

Miles, M.B., &Huberman, A.M. (1994). *Qualitative data analysis*. London: Sage.

Parasuraman, A., Berry, L., Zeithaml, & Valarie, A. (1985). A conceptual model of service quality and its implications for future research. *Journal of Marketing*, *49*(4), 41-50.

Pilling, J. (1999). Problems facing the radiologist tendering for a hospital wide PACS system. *European Journal of Radiology, 32*, 101–105.

Protopapas, M., Lappas, D., Keitel, H. & Knight, P. (1996) Recollisions, bremsstrahlung, and attosecond pulses from intense laser fields. *Physical Review online archive*, *53*(5).

Reiner, B., Siegel, E., Charles, F., Hooper, F., Cox, R. & Scanlon, M. (2000) Effect of filmless imaging on the utilization of radiology services. *Radiology, 2* (14), 163-167

Reiner, B.I., & Seigel, E.L (2002). Application service providers: An alternative approach to PACS implementation. *Journal of Digital Imaging, 14*, 1–8.

Rostenberg, B., (2006) *The architecture of medical imaging*. Hoboken, NJ: John Wiley & Sons, Inc.

Rumreich, L. & Johnston, A. (2003) From traditional reading rooms to a soft copy envirnoment: Radiologist satisfaction survey. *Journal of Digital Imaging, 16*(3), 262-269.

Scott, W. et al. (1995). Interpretation of emergency department radiographs by radiologists and emergency medicine physicians: Teleradiology workstation versus radiograph readings. *Radiology, 195*, 223–229.

Seddon, P. B. (1997) A respecification and extension of the DeLone and McLean model of IS success. *Information System*, *8*(3), 240-253.

Slasky, B., et al. (1990). Receiver operating characteristic analysis of chest image interpretation with conventional, laser-printed, and high-resolution workstation images. *Radiology, 174*, 775–780.

Speier, C., & Venkatesh, V. (2002). The hidden minefields in the adoption of sales force automation technologies. *Journal of Marketing*, *66*(3), 98-111.

Strauss, A. (1987). *Qualitative analysis for social scientists*. Cambridge: Cambridge University Press.

Ulrich, D. (1991). Using human resources for competitive advantage. In I. Kilmann (Ed.), *Making organizations competitive*. San Francisco: Jossey-Bass.

Watkins, J., Weatherburn, G., & Bryan, S. (2000). The impact of a PACS upon an intensive care unit. *European Journal of Radiology, 34*, 3–8.

KEY TERMS

Environmental Training: Represents activities that teach staff how to better perform their present jobs in order to cope with environmental changes, such as changes in lighting, monitors, and so forth.

Evaluation: The systematic acquisition and assessment of information using an overarching perspective to provide useful feedback about some sort of object/subject.

Hospital Strategy: An integrated set of actions with a focus on increasing the long-term prosperity and strength of hospitals relative to its environment and (strategic) goals.

Information Quality: Concerned with such issues as the relevance, timeliness, and accuracy of information generated by an information system (Seddon, 1997).

Information Systems: A system that comprises people, machines, and/or methods organized to collect, process, transmit, and disseminate data that represent information.

Not for Profit: The concept that the primary objective is to support some issue or matter of private interest or public concern for noncommercial purposes.

Service Quality: The degree of inconsistency between customers' normative expectations for the service and their perceptions of service performance (Parasuraman et al., 1985).

Strategy: A pattern of decision in a company that determines and reveals its objectives, purposes, or goals; produces the principal policies and plans for achieving those goals; and defines the range of business the company is to pursue, the kind of economic and human organization it is or intends to be, and the nature of the economic and noneconomic contribution it intends to make to its shareholders, employees, customers, and communities.

System Quality: Concerned with whether or not there are "bugs" in the system, the consistency of the user interface, ease of use, quality of documentation, and sometimes quality and maintainability of the program code (Seddon, 1997).

Viewing Environment: The characteristics of the workspace in which digital images are viewed.

Theory Driven Organizational Metrics

W.F. Lawless
Paine College, USA

Joseph Wood, LTC
Fort Gordon, USA

Hui-Lien Tung
Paine College, USA

INTRODUCTION

The inability to establish first principles has kept organizational theory from being successful. Moreover, due to snapshots in time and researcher biases, case studies are limited to hindsight, rather than serving as a proactive source of solutions to organizational problems. Yet case studies guided by theory have illuminated and tested the first principles that we have discovered. Unlike simple Newtonian mechanics, however, social-psychological mechanics among organizational members are hidden behind and within explanations and discourse, eluding a science of fundamental interactions. When an interaction stops for measurement (e.g., case studies), significant information from the collapse of organizational interdependence is lost. The path forward is to predict the uncertainty left from the collapse of interdependent variables: planning and execution; or resources and time. In this article, we develop a new organization theory; in a related article ("Restructering a Military Medical Department Research Center" in this encyclopedia), we apply the theory to a case study of a military medical research center (MDRC) with access to advanced information systems (IS), yet struggling to determine the quality of its residents in training, and their scholarly productivity.

BACKGROUND

Organizational theory has failed to produce predictable (Pfeffer & Fong, 2005) or replicable results (Weick & Quinn, 1999). Traditional organizational theory, generally based on (MI) (e.g., game theory; in Nowak & Sigmund, 2004), assumes that open information from individuals is stable and freely accessible (Baumeister, Campbell, Krueger, & Vohs, 2005), making an organization a rational aggregation of the contributions from its individual members. Yet defining rationality as "normative consistency," Shafir and LeBoeuf (2002) concluded that neither average humans nor experts make consistent choices, preferences, or justifications, undercutting the traditional model of rationality. Agent-based models (ABMs) based on MI with the same assumption about rational aggregation have had to discount the value of prediction as a consequence: "the value of a [computational] model is not prediction but insight" (Bankes, 2006). But to successfully operate an autonomous computational organization in the field (robot teams, human-robot teams, and networked and virtual organizations), a rational process using information systems (IS) and technology (IT) to formulate predictions is necessary to protect humans, the environment, and to ensure the execution of missions.

THEORY AND ORGANIZATIONAL METRICS

In contrast to methodogical individualism (MI), adopting the quantum uncertainty relations as first suggested by Bohr (1955) and Heisenberg (1958), has begun to successfully model interdependent uncertainties in human social interaction (Lawless, Castelao, & Abubucker, 2000) to predict decision-making among human organizations in the field (Lawless, Bergman, & Feltovich, 2005), and to study organizations in the laboratory (Lawless, Bergman, Louca, & Kriegel, 2006b). Subsequently, online metrics based on the "measurement paradox" have been proposed (Lawless & Grayson, 2004).

The paradox indicates that measuring an interaction or organization collapses the existence of its interdependent information into strictly classical information that cannot be aggregated to reconstruct the organization (Levine & Moreland, 1998), nor apparently even to reconcile differences between individual beliefs and actions—despite more than 30 years of research, no better than a weak link has been confirmed between selfesteem and actual performance at school or in the workplace (Baumeister et al., 2005). In addition, it has been known for some time that surveys or case study interviews fail to generate information that can validly predict individual and organizational change (Eagly & Chaiken, 1993). But surprisingly, the measurement paradox suggests that the collapse of interdependent information can be exploited to favor one of two interdependent states in the mathematical model of interdependence to produce predictable outcomes under certain rather extreme conditions, such as the difference between consensus (CR) versus majority rule (MR) decision processes in organizations (Lawless et al., 2005): It has been predicted and found that CR leads to less concrete decisions less welcomed by an organization's customers, but at lower resource expenditures that take longer to process; in contrast, MR leads to more practical decisions more welcomed by customers, but with more conflict and resources expended that quicken decisions.

The development of a complete theory of organizations, however, requires concepts beyond extreme situations to more general cases. Recognizing that two stories are always possible (Wendt, 2005), the measurement paradox arises because no better than one story at a time is ever collected during measurement. Further, the relationship between decision processes and organizations is itself complex, especially for CR. The purpose of CR is to convert the neutrals in a group into active individual participants (Bradbury, Branch, & Malone, 2003). However, the process in CR that suspends the criticism of beliefs no matter how bizarre lends itself to being hijacked: "The requirement for consensus in the European Council often holds policy-making hostage to national interests in areas which Council should decide by a qualified majority" (WP, 2001, pp. 29). Organizations are primarily hierarchical and governed by a single leader or command decision-making (CDM), making the link to CR more obvious under the control of multiple leaders (e.g., the leadership crisis at Unilever prior to 2005; the current management crisis at Europe's aerospace EADS group that includes Airbus). However, single leaders using intimidation or even violence can convert an organization or system into a quasi-CR process that stifles widespread criticism; for example, Germany's lack of response in 1934 from its citizens or institutions to the multiple murders during Hitler's "Night of the Long Knives" (Benz, 2006, pp. 54). But, counter-intuitively, instead of actively seeking consensus (CR), it has been found that the most robust consensuses are derived during competitive decision-making (i.e., MR), more learning occurs under competition (Dietz, Ostrom, & Stern, 2003), and the more competitive is a team, the greater the cooperation among its members (Lawless et al., 2000)[1].

In addition to laboratory studies, the paradox has been exploited by proposing the first mathematical set of interdependent metrics designed to measure the real-time performance for a system of military forecasters in the field (Lawless, Bergman, & Feltovich, 2006a). These metrics were revised and extended to analyze the reorganization of IT services provided by the Management Information Service Center (MISC) at a major university in Europe to test first principles that were then used to reverse model terrorist organizations (Lawless et al., 2006b). Mindful that a case study reflects a static snapshot in time which exposes the findings from a case study to confirmation bias (Eagly & Chaiken, 1993), the potential for biases were countered with a theoretical foundation directed at two sets of two interdependent variables (in Figure 1, planning and execution; resources and time). Based on these four interdependent variables, it was found that corruption at MISC and its university occurred by operating without a structured business model (BM). The lack of a focused BM for the university had led to a disorganized assemblage of faculty, staff, and students that discouraged innovation, promoted administrative malfeasance, and resource mismanagement, impeded student progress and faculty research, and significantly reduced opportunities for MISC and its university. The conclusion followed that a loose aggregation in the limit approaches a CR process, in that less information is processed by the organization than its members, consequently precluding organizational learning and change in response to environmental perturbations (Dietz et al., 2003). The results helped to specify a computational model of an organization with artificial agents that could be used as a test laboratory for organizations, and also used as online metrics for traditional and virtual organizations, based on applied IT algorithms.

In Figure 1, uncertainty in the social interaction is represented by an interdependence between strategy, plans, or knowledge uncertainty, ΔK (where K is a function of the social location where it was learned; from Latané, 1981), and uncertainty in the rate of change in knowledge or its execution as $\Delta v = \Delta\ (\Delta K/\Delta t)$; this agrees with Levine and Moreland (2004), that as consensus for a concrete plan increases (ΔK reduces), the ability to execute the plan increases (Δv increases). By extension, interdependence also exists in the uncertainty in the resources expended to gain knowledge, ΔR, and by uncertainty in the time it takes to enact knowledge, Δt (for proofs, see Lawless et al., 2006b). That these two sets of bi-stable factors are interdependent means that a simultaneous exact knowledge of the two factors in either set is precluded.

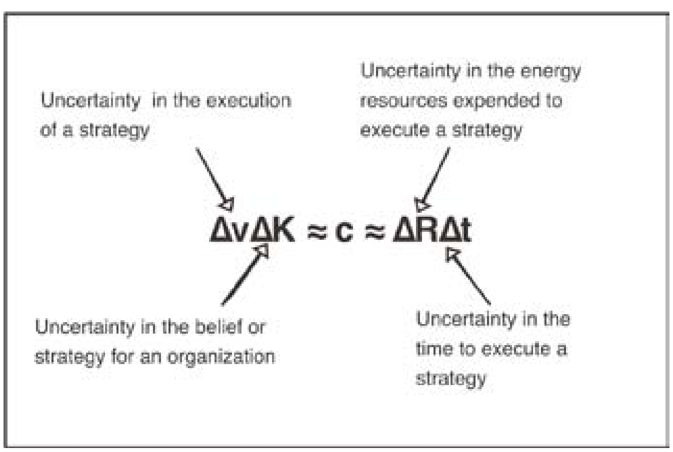
CASE STUDY: APPLICATION OF PRINCIPLES TO MDRC

In 2006, briefly, a new director assumed command of a military Medical Department Research Center (MDRC) at a regional military medical hospital. MDRC supports clinical and basic research for its staff and all hospital personnel, including family practitioners, internal medicine, general and orthopaedic surgery, and dentistry. In addition to providing basic research support for the hospital, MDRC is responsible to teach the fundamentals of experimental research to the hospital's medical residents[2], to provide continuing education for more experienced care providers, and to train dentists in research methods. One of the goals of the training by MDRC is to help the hospital's Graduate Medical Education (GME) candidates become certified by their respective American specialty boards, in what is becoming a matriculation requirement for many training programs.

The Chief, MDRC, wants to establish Metrics of Effectiveness (MOE) to measure his organization's success, and to craft a plan with MOEs to improve the performance of scholarly activities (i.e., Business Model (BM)). One problem with using MOE's is that much of the research proposed in the MDRC protocols has lasted or may last for a number of years before scholarly products can be published, whereas other research protocols may last under one year, giving the less complex Protocols an advantage in the generation of scholarly products. As the complexity of a Protocol increases, the time necessary to complete a program of research also increases. The Chief wants to increase the complexity of protocols, but remain able to measure the impact of complexity on scholarly productivity[3].

Fragmentation from a lack of focus indicates a general reduction in motivation to serve an organization, and consequently the need to change the business model to refocus the organization. The present case study of MDRC indicated that fragmentation among its processes, and researchers had reduced control over future plans. On the one hand, fragmentation of an organization is associated with increased innovation at the individual level (e.g., Benz, 2006), but often at the

Figure 1.Organizational uncertainty



expense of the organization. Successful organizations are entangled by command decision-making (CDM) into centers of cooperation (Lawless et al., 2006a; Lawless et al., 2000b) to acquire new resources and reduce wasteful expenditures, but that also tend to reduce innovation from their marginalization of new knowledge in making the organization, in this case MDRC, more competitive in its marketplace for new grants, improved mentors, better trainee candidates, and new joint ventures with industry or other research facilities. But this reduced innovativeness can be countered by mindfully taking steps with IT to free MDRC from many labor intensive activities by automating data collection and analytical processes in order to free resources to increase innovation. The overall goal of implementing the recommendations is to reduce the overall costs of MDRC, increase its operational efficiency, and to increase its growth by enhancing its ability to attract new extramural funds.

CASE STUDY: MEASUREMENT RESULTS

The results from data collection at MDRC invokes the measurement problem for the four interdependent factors in Figure 1.

Planning Uncertainty for the Organization's Business Model (BM, or ΔK)

In the MDRC Chief's opinion, attracting better quality residents will require an increase in the quality of publications, and in receiving help from other Division Chiefs, plus having a stronger program. Then better graduates will begin to attract other quality candidates, adding credibility to the BM.

The MDRC Chief observed that the support provided by his organization is satisfactory from all accounts, but that it should be better organized and more focused (INC, 2006). The revised BM should permit everyone to work under the same roof, gain more extramural funds, improve scholarly productivity, and keep education as the primary goal of MDRC. As part of an innovation plan, new resources or funds must be sought from State, Federal, and industrial sources. If successful, the plan should produce a qualitative shift among mentors and GME trainees under the new system; trainees will also learn new professional techniques that should help them to find better jobs afterwards.

The new BM must focus organizational resources to produce high quality research executed under a sense of competitive urgency. The goal should not be to seek complexity, but instead publishable research that helps to drive more and more competitive scholarly productivity, and the search for extramural funds. Complexity should be a byproduct of the plan.

Planning Execution (Δv)

The goal of executing the BM should be to increase the number of customers per period (new grant funds; new trainees; and new recognitions of quality).

The lack of focus at MDRC has primarily arisen from an internal lack of competition among its organization's members in the execution of its current BM. However, a lack of focus again suggests that an implicit CR exists to block the execution of a revised BM. Resistance to implementing a new BM can be anticipated, implying that barriers must be anticipated and overcome, as well as an average rise in activation that should be exploited to increase the rate of execution.

With support from his fellow Division Chiefs, colleagues, and staff, eventually the barriers that arise can be overcome among new mentors and new trainees around which time a new culture should be encouraged to become established and grow.

The MDRC and hospital staff should be educated to understand the need for regular professional training to improve research performance, especially the quality and the quantity of scholarly productivity; and the need for new information channels to distribute technical information about research opportunities. Numerous messages about the change should be given to trainees in seminars provided by MDRC, and designed to revise the culture to match the new BM.

Resource Uncertainty (ΔR)

The goal of the plan of action (BM) is to maximize the resources available to MDRC to execute its BM in the minimum of time. Such a BM, however, will likely reduce innovation, but practices instituted to seek innovation can offset this shortfall (e.g., including in the BM a strategy to continually seek new partners to obtain extramural grants, along with MOE's of grant progress).

One measure of progress is the level of teamwork across the organization (MDRC staff, mentors and trainees) to increase the competitiveness of MDRC's performance in completing its protocols as an indirect measure of the overall effort by its research teams; however, teamwork must not be simply commanded, but encouraged as a part of the competitive process and demonstrated to work.

As innovation increases from the gathering and expenditure of resources derived from the discovery of new resources (new trainees and new faculty become more attracted by new skills and the initiation of new industrial-state-federal projects), and a reduction of costs (less waste), planning complexity increases correspondingly as the ability to direct these freed funds to a random exploration of new projects for MDRC and the hospital with success determined by a reduced effort (practice effects; stochastic resonance) in discovering new sources of free energy.

Time Uncertainty (∆t)

The goal is to reduce the time to execute a BM, gain new resources, and discover new opportunities.

As the available resources are increasingly directed by MDRC to the completion of existing projects with the ultimate goal of freeing resources for new projects, the average time to complete and execute existing projects should decrease. At the same time, if and only if an innovation circuit has been established, the time to innovate should decrease correspondingly as new opportunities arise.

As new opportunities become available and exploited by MDRC, they will provide new opportunities to trainees, adding incentives that should further improve the quality of future trainee candidates.

FUTURE TRENDS

Should our initial exploitation of the measurement paradox for MDRC prove fruitful in the design of new organizational metrics, we expect to shift from the static metrics we have envisioned in this study to dynamic and online metrics that MDRC can deploy in real-time to tune and execute the strategic vision in its Business Model (BM). MDRC will do this with feedback derived by focusing its BM on its core missions, by optimally aligning its scarce resources to drive its vision forward, by reducing existing barriers to its strategy, and by seeking opportunities to exploit its strategy, as well as new opportunities to employ its organizational expertise.

We also expect to create a model of MDRC with artificial agents that could be used as a test laboratory to explore variations in its strategy, variations in its execution, different sources of new resources and barriers, and variations in the time it takes to execute its BM and to deploy its resources.

We have also recently begun conversations with other MDRC-like units in the military. We are exploring the possibility of an online metric for the entire system to reduce the costs of automating these metrics. It will consist of a system-wide automated electronic Protocol System (ePS), an automated Document Management System (DMS) for academic manuscripts, and an automated Metrics System (for a source of the value of an electronic Protocol-IRB System, see Hood, Gugerty, Levine, & Ho, 2005).

CONCLUSION

MDRC is located at one of the DoD's most advanced IS bases, yet its administrators have been using ad hoc processes to manage its protocols, educate its residents, improve patient care, and track manuscripts. Guided by theory, we hence exploited the measurement problem to construct a set of metrics to help MDRC restructure its operation and measure its performance. The resulting metrics represent a significant step forward in applied organizational theory.

REFERENCES

Bankes, S.C. (2006, June 23). *Keynote*, NAACSOS-06, Notre Dame.

Baumeister, R.F., Campbell, J.D., Krueger, J.I., & Vohs, K.D. (2005, January). Exploding the self-esteem myth. *Scientific American*.

Benz, W. (2006). *A concise history of the Third Reich*. Dunlap, T. (Trans.). Berkeley, CA: UCLA Press.

Bohr, N. (1955). Science and the unity of knowledge. In L. Leary (Ed.), *The unity of knowledge* (pp. 44–62). New York, Doubleday.

Bradbury, J.A., Branch, K.M., & Malone, E.L. (2003). *An evaluation of DOE-EM public participation programs (PNNL-14200).* Richland, WA: PNNL.

Dietz, T., Ostrom, E., & Stern, P.C. (2003). To govern the commons. *Science, 302,* 1907.

Eagly, A.H., & Chaiken, S. (1993). *The psychology of attitudes*. Fort Worth, TX: Harcourt.

Hardin, G. (1968). The tragedy of the commons. *Science, 162,* 1243–1248.

Heisenberg, W. (1958/1999). Language and reality in modern physics. In *Physics and Philosophy. The Revolution in Modern Science* (pp. 167–186). New York: Prometheus Books.

Hood, M.N., Gugerty, B., Levine, R., & Ho, V.B. (2005) Computerized information management for institutional review boards *Computers, Informatics, Nursing, 23*(4), 190–198.

INC. (2006). Interview notes dated June 2, 2006, and subsequent informal communications.

Kruglanski, A.W., Pierro, A., Mannetti, L., & De Grada, E. (2006). Groups as epistemic providers: Need for closure and the unfolding of group-centrism. *Psychological Review, 113,* 84–100.

Latante, B. (1981). The psychology of social impact. *American Psychologist, 36,* 343-356.

Lawless, W.F., Bergman, M., & Feltovich, N. (2005). Consensus-seeking versus truth-seeking. *ASCE Practice Periodical of Hazardous, Toxic, and Radioactive Waste Management, 9*(1), 59–70.

Lawless, W.F., Bergman, M., & Feltovich, N. (2006a). A physics of organizational uncertainty. Perturbations, measurement and computational agents. In Shu-Heng Chen, Lakhmi Jain, & Chung-Ching Tai (Eds.), *Computational Economics* (pp. 268–289). Hershey, PA: Idea Group Inc.

Lawless, W.F., Bergman, M., Louca, J., & Kriegel, N.N. (2006b, forthcoming). *A quantum metric of organizational performance: Terrorism and counterterrorism, computational and mathematical organization theory*. Springer.

Lawless, W.F., Castelao, T., & Abubucker, C.P. (2000). Conflict as a heuristic in the development of an interaction mechanics. In C. Tessier, L. Chaudron, & H. Muller (Eds.), *Conflicting Agents: Conflict Management in Multi-Agent Systems*. Boston: Kluwer.

Lawless, W.F., & Grayson, J.M. (2004). A quantum perturbation model (QPM) of knowledge and organizational mergers. In L. van Elst, & V. Dignum (Eds.), *Agent Mediated Knowledge Management* (pp. 143–161). Berlin: Springer.

Levine, J.M., & Moreland, R.L. (1998). Small groups. In Gilbert et al. (Eds.), *Handbook of Social Psychology* (pp. 415–469). McGraw.

Nowak, M.A., & Sigmund, K. (2004). Evolutionary dynamics. *Science, 303*, 793.

Pfeffer, J., & Fong, C.T. (2005). Building organization theory from first principles. *Organization Science, 16*, 372–388.

Shafir, E., & LeBoeuf, R.A. (2002). Rationality. *Annual Review of Psychology, 53*, 491–517.

Smith, W.K., & Tushman, M.L. (2005). Managing strategic contradictions: A top management model for managing innovation streams. *Organizational Science, 16*(5), 522–536.

Weick, K.E., & Quinn, R.E. (1999). Organizational change. *Annual Review of Psychology, 50*, 361–386.

Wendt, A. (2005). Social theory as Cartesian science: An auto-critique from a quantum perspective. In S. Guzzini, & A. Leander (Eds.), *Constructivism and International Relations* (pp. 181–219). New York: Routledge.

Wood, J. (2004). *MDRC Clinical Investigation Program Annual Report 2004, RCS MED-300 (R)*. For access, contact joseph.c.wood@us.army.mil

Wood, J. (2005). *MDRC Clinical Investigation Program Annual Report 2004, RCS MED-300 (R)*. For access, contact joseph.c.wood@us.army.mil

WP. (2001). White Paper. European governance (COM (2001) 428 final; Brussels, 25.7.2001).

KEY TERMS

Agent-Based Model (ABM): Used to create complex models of organizations and systems; for example, with Monte Carlo methods, ABM's can be designed to reduce biases in Critical Path Method (CPM models and Program Evaluation Review Technique (PERT)) models.

Business Models (BM): A business model is a complex plan for business operations and strategy, that when it converges into a consensus, it maximizes productivity, but the stronger the consensus in support of the BM, the more that innovation is reduced. Alternatively, divergence increases opportunities for innovation, but by reducing productivity. This paradox is reduced by creating ambidextrous organizations that can operate in a state of tension between both increasing productivity and innovativeness (Smith & Tushman, 2005).

Command Decision-Making (CDM): Top-down decision-making, especially the autocratic decision-making practiced in industry, business, and the military, but also by dictators. CDM reduces innovativeness but increases productivity. Further, CDM often employs consensus rules in its decision-making processes, because CR is open to exploitation (Kruglanski, Pierro, Mannetti, & De Grada, 2006).

Consensus-Seeking Rules (CR): By reducing evidentiary barriers to discussion, consensus-seeking rules often lead to the least common worldview amongst risk perceivers, making it unlikely to reach a practical decision, other than to instantiate past organizational practices, but consequently, reducing innovativeness. Further, consensus rules require considerable time to listen to each risk perception.

Majority Rules (MR): Majority rules often lead to faster decisions, more practical decisions, and, counterintuitively, stronger consensuses. The problem with majority rules is that they introduce conflict into decision-making. But if the conflict can be moderated or managed, the result is more learning among the participants compared to consensus rules.

Measurement Problem/Paradox: Assuming the existence of uncertainty in the four interdependent variables of planning-execution and resources-time, decreasing the uncertainty in either set raises the uncertainty in its correspondingly linked interdependent variable.

Methodological Individualism (MI): Game theory (Nowak & Sigmund, 2004) has attempted to substantiate the superiority of cooperation based on the belief that cooperation leads to the highest social good, even if it is coerced (Hardin, 1968). However, the lack of substantiating evidence in support of the social worth of cooperation is reviewed in Lawless and Grayson (2004).

Metrics: The metrics for MDRC are being designed to convert the measurement problem into an organizational metric of performance.

ENDNOTES

1 We speculate that the consensus produced by MR is more action oriented than the world view consensus produced by CR, and, therefore, qualitatively different.

2 Categorical residents working within a specialty, residents rotating among the hospital's different specialties, and transitional interns; in JWO, 2005, pp. 4.

3 For the entire case study, see case study article (Lawless et al., 2007).

Time-Frequency Analysis for EGM Rhythm Classification

Hamid Sheikhzadeh
ON Semiconductor, Canada

Robert L. Brennan
ON Semiconductor, Canada

INTRODUCTION

In this article, we study the problem of rhythm classification and event detection based on intracardiac electrogram (EGM) signals. At present, only a very limited scope of signal processing is possible in implantable cardioverter defibrillators (ICDs), due to the scarcity of available resources. As a result, relatively simple beat-by-beat time-domain analysis of the EGM signal(s) is employed for rhythm detection. Recently, researchers have attempted to exploit more sophisticated signal processing methods, such as wavelet transforms and template matching (Astrom, Olmos, & Sornmo, 2006; Brown, Christensen, & Gillberg, 2002; Koyrakh, Gillberg, & Wood, 1999). However, the new methods have rarely been employed in practical systems because of their computational and power demands.

Modern implantable cardiac rhythm management (CRM) systems such as ICDs face many challenges, some of which are summarized here:

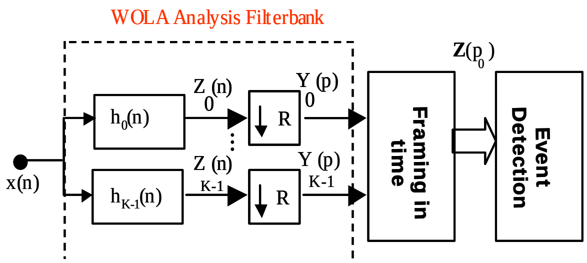
1. Quick and reliable detection of serious cardiac events is yet to be achieved by CRM devices. Errors in event detection occur, due to many reasons, such as quick morphology, rate, and polarity changes of the EGM signal, abnormally wide R-waves and P-waves, and external noise (Astrom et al., 2006).
2. Modern CRM devices are challenged by a considerable rate of inappropriate device therapy (IDT) (between 10 to 30%) (Rojo-Alvarez, 2002). Major causes of IDTs include: poor EGM signal quality, sinus tachycardia, supraventricular tachycardia (SVT), myopotential interference, external interference, and T-wave oversensing (Schaer, 2000). In addition to being painful for the patient, IDT's deplete valuable device battery power, and potentially place the patient at serious risk (Rojo-Alvarez, 2002; Schaer, 2000).
3. Rapid redetection of the EGM rhythm after device therapy (low-energy pacing or high-energy cardioversion) is problematic for CRM devices. This is particularly difficult when utilizing time-domain methods, due to their sensitivity to polarization of the EGM electrodes or baseline variations of the EGM signal following device therapy.
4. Modern CRM devices offer increasingly sophisticated multichamber detection and control, offering the physicians more potential therapies. In absence of powerful signal processing methods, however, the physicians are unable to fully exploit the multichannel information for effective programming of the device. The multichamber detection systems also require increased immunity to crosschannel interference.

In view of these challenges, we propose a multitiered detection method offering improved rhythm detection accuracy for CRM devices. This method, based on time-frequency analysis of the EGM signals, is targeted for implantable devices operating with extremely low-power budgets. It operates in real-time, and the processing delay is in the acceptable range for these applications. Although the method employs a specific filterbank implementation, it is not limited to this specific filterbank. Instead, the point is to prove that parallel processing of subband signals, resulting from a time-frequency analysis, offers more reliable detection in compromised situations.

BACKGROUND

Time-frequency analysis has already been exploited to extract features for electrocardiogram (ECG) rate and QRS detection (Kohler, 2002).

Figure 1. WOLA oversampled filterbank analysis for subband rhythm detection



Among various time-frequency analysis methods, oversampled filterbanks have been implemented by a Weighted Overlap-Add (WOLA) structure in ultra-low resource devices (Brennan & Schneider, 1998). We, therefore, propose a cardiac event detection based on the WOLA analysis of the EGM signals, and demonstrate its capabilities in various signal conditions. As the intention here is to describe the basis for the detection method, we limit our attention to single-electrode analysis; an extension to multiple-electrode analysis is straightforward.

After surveying the literature, the contribution by Afonso et al. (1999) on ECG beat detection in real-time is most relevant to our research. Afonso analyzes the ECG signal by a critically sampled polyphase filterbank to extract six features that are all based on the energy (or absolute value) of groups of subbands to obtain six time-frequency signals. Each of these signals is then passed through a peak detector that compares the signal moving average to a threshold that is determined based on an estimation of the background noise. Next, in a cascade of five stages (levels), peaks are further refined. This method was evaluated on a standard ECG database with satisfactory performance.

Afonso's method and similar methods proposed for ECG analysis are, unfortunately, not applicable to ultra-low power implantable CRM devices using EGM signals for the following reasons:

1. The sensed EGM signals and ECG signals differ in many aspects.
   a. ECG signals provide a composite signal after propagation of various electric waves to the body surface. In contrast, EGM signals provide direct access to individual heart chambers at the signal source.
   b. Relative timing of various EGM signals is crucial in discrimination of various cardiac events. For ECG signals however, such timing information is not available.
   c. Unlike ECG signals, sensed EGM signals are prone to cross-talk.
2. The ECG methods are too complex or unsuitable for real-time applications, are specifically designed for (and evaluated on) ECG signals, and do not provide the robust and reliable performance essential for EGM signal processing.

SUBBAND-BASED CARDIAC EVENT DETECTION

Overview

A time-domain digital EGM signal $x(n)$ is analyzed by an oversampled filterbank (depicted in Figure 1) with K subbands. For efficiency, a WOLA structure is adapted that can be implemented with ultra-low resources (Brennan & Schneider, 1998, 2003). The filterbank parameters are: $K=32$ subbands, analysis filters $h_k(n)$, $k= 0, 1, \ldots, K$-1, each of length of $L=256$, subband decimation of $R=4$, and oversampling factor of $OS=K/R=8$. These parameters were optimally selected for the EGM signal analysis through experimentation. As depicted in Figure 1, the subband signals are decimated with respect to the EGM signal. Figure 1 only depicts a schematic representation of the WOLA analysis. The actual WOLA implementation based on Brennan and Schneider (1998, 2003) is involved, and not described here. The WOLA time-frequency analysis stage produces K complex-valued subband signals $Y_k(p), k= 0, 1, \ldots, K-1$, where $Y_k(p)= Z_k(n.R)$, and k is the subband time-index. Due to Hermitian symmetry, half of the subbands are unique for real signals, because the other half are complex conjugates. For the next stage, subband signals are framed with blocks of M samples (in time) for further analysis. For a subband time-index of p_0, this generates a $K/2$-by-M matrix $\mathbf{Z}(p_0)$ including subband signals $Y_k(p), k= 0, 1, \ldots, K/2-1, p=p_0-M+1, (p_0-M+2),..,p_0$. The frame-length M and the frame-shift p_0 were selected to be three and two seconds, respectively. The frame-length should be long enough to cover more than one beat for slow beats (around 60 beats per minute (bpm)) for statistical analysis. At the same time, the frame should be as short as possible to track the dynamics of potentially quickly varying beats. The choice of frame-shift is rather arbitrary, and depends on how often a decision is needed. Notice that irrespective of the frame-length and frame-shift, the WOLA analysis is continuously applied to the input signal, yielding a new block of subband signals every R input samples.

The basic idea behind the proposed subband detection method is the concept of "subband synchrony" described now. Since the cardiac beat generates a sharp pulse, the magnitudes of subband signals ($|Y_k(p)|$) exhibit mainly coherent peaks at the time of cardiac depolarization. By properly exploiting this subband coherence (or "synchrony") in selected subbands, one can achieve robust peak detection. We propose a simple detection method based on binary operations to exploit the subband synchrony. Of course, various other methods might be designed based on the same principle.

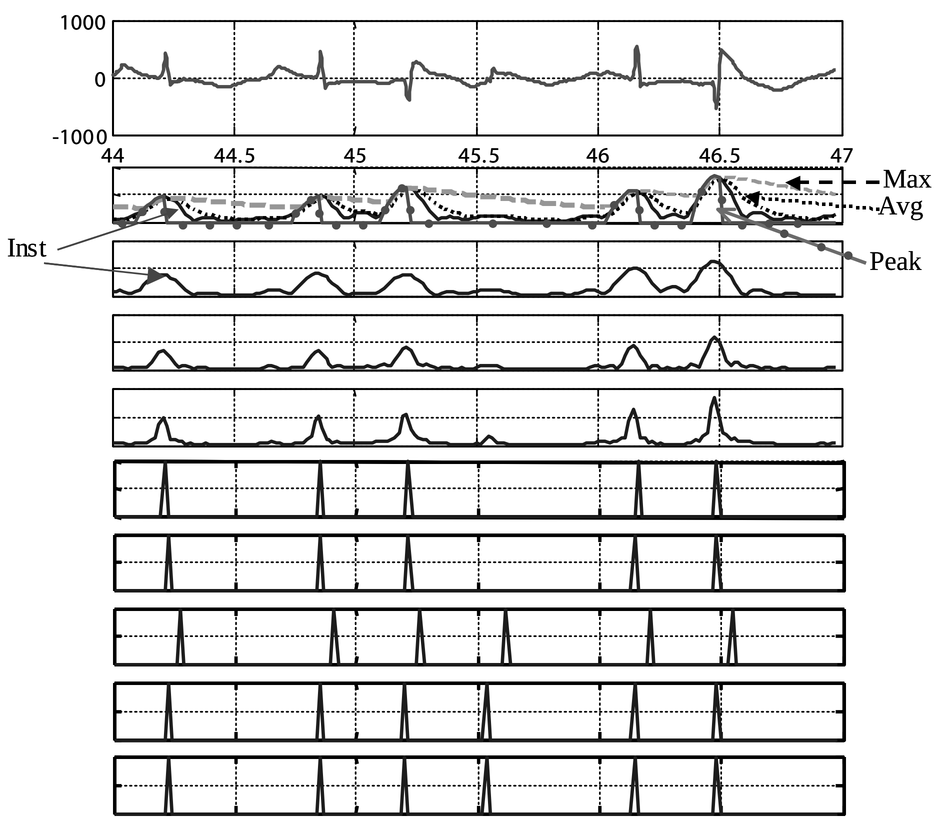
After the synchrony analysis, subband beats are combined to obtain a "final beat" sequence for every $\mathbf{Z}(p_0)$ matrix. Then the periodicity and the regularity of the final beat, together with the synchrony measure, are considered for event detection.

Subband Peak and Synchrony Detection

Each subband magnitude signal $|Y_k(p)|$ in signal matrix $\mathbf{Z}(p_0)$ is tracked in time to detect its peaks. In each subband, we tracked the maximum with a two-time-constant first-order recursive filter. For a subband magnitude signal $|Y_k(p)|$, two time-constants of $0.9<\alpha_{m1}<1$ and $0.1<\alpha_{m2}<0.5$ are employed to obtain the maximum signal $M_k(p)$ as follows:

$$\alpha=\alpha_{m1}$$
$$\text{If } |Y_k(p)| > M_k(p-1),\ \alpha=\alpha_{m2}$$
$$M_k(p) = M_k(p-1).\alpha + |Z_k(p)|.(1-\alpha)$$

Figure 2. From top to bottom, A segment of ventricular EGM signal, four WOLA subband energies (with average, maximum, and peak signals sown for the first subband), their corresponding binary pulses $B_k(p)$ (rows 6-9), and the final detected pulse (bottom row)



After each subband peak, the filter acts as a leaky integrator. Thus, the first filter coefficient (α_{m1}) is selected close to one to control the so-called "release-time" of the filter. The other smaller coefficient α_{m2} allows the filter to react quickly to the next peak.

A similar average tracking with a two-time-constant recursive filter is used to track the average signal ($A_k(p)$) with $0.7<\alpha_{a1}<1$ and $0.5<\alpha_{a2}<0.7$. A peak ($P_k(p)$) is detected by comparing the instantaneous value ($|Y_k(p)|$) to the average and maximum values, as described in the following pseudo-code:

$P_k(p)=0$
If $\{|Y_k(m)| > A_k(p\text{-}1)$ & $|Z_k(p)|> 0.5\ M_k(p\text{-}1)\}$ &
  $\{A_k(p\text{-}1)< 0.9\ M_k(p\text{-}1)\}$,
$P_k(p)= |Y_k(p)|$

In absence of EGM activity, $P_k(p)$ and $M_k(p)$ converge into each other. In this situation, the last term in the condition prevents peak detection. Currently, peak detection is limited to subbands 2-9, ignoring the first subband, as it is compromised through the capture of noise and baseline wander. Also high-frequency subbands ($k>9$) were ignored, due to low levels of synchrony in them. Figure 2 depicts a typical segment of ventricular EGM signal (top graph), and four WOLA subband instantaneous energies with maximum, average, and peak signals shown for one of the subbands.

To simplify further operations, we convert each subband peak signal $P_p(p)$ to a binary (0/1 for peak/no-peak) signal $B_k(p)$ after peak detection. Once a pattern of consecutive peaks followed by zeros (e.g., {1 1 1 0 0}) is found, the falling edge of the peak is registered as a valid peak. The rest of the peak signal is reset to zero. This avoids detecting short-term spurious peaks. All further operations are applied to the binary peak signals $B_k(m)$. Next, the degree of synchrony between various subband peaks is measured by simple AND operations of binary peak signals. Given a pair of signals $B_k(p)$ and $B_l(p)$, synchrony of the pair $S_{k,l}$ is measured as the ratio of coinciding peaks in the pair over the maximum number of peaks in the two signals.

The synchrony of all nonidentical pairs is evaluated, and the top three synchrony scores are kept as measures of the frame synchrony. Based on the top 3 scores, using fixed synchrony thresholds, the detection algorithm classifies the subband beats as perfectly synchronous (*Syn=4*), as borderline synchronous (*Syn=2*), or asynchronous (*Syn=0*).

The signals in each of the top three most synchronous binary pulse pairs are combined in pairs by the AND operation. Next, the results are used for robust detection of beat times. Applying a majority voting rule (two out of three) and using the AND results, the final beat times are detected. This proved to be very robust in noise, and especially when the signal quality was compromised in flutter and fibrillation. The result of beat detection is a final beat sequence per frame. Figure 2 (rows 6-9) depicts binary subband pulses for the EGM segment, together with the final detected beat (last row).

Periodicity and Regularity Analysis

The detected beat signal is further analyzed to find its period histogram, its mean period ($\bar{T}$), and the mode of the histogram (T_m). Also, the standard-deviation-to-mean ratio of the detected periods (σ/μ) is calculated. σ/μ values more than 40% indicate very irregular beats, values less than 20% show very regular beat patterns, and values in between indicate moderate regularity. Also, if unusual lack of EGM activity is detected within a frame, it indicates irregularity.

Cardiac Even Detection

Based on the synchrony score (*Syn*), periodicity ($\bar{T}$ T_m, σ/μ), and regularity of the final beat sequence, cardiac events are classified to be one of the following 8 events:

1. Stable Sinus Rhythm (SR)
2. Transitional SR (T-SR)
3. Stable Tachycardia (VT or AT)
4. Transitional Tachycardia (T-VT or T-AT)
5. Flutter (VFLUT or AFLUT)
6. Fibrillation (VFIB or AFIB)
7. Synchronous but Irregular rhythm (Syn-Irg)
8. Unclassified

Represented by VT, VFLUT, and VFIB are ventricular events of tachycardia, flutter, and fibrillation respectively. Similarly, AT, AFLUT, and AFIB represent the corresponding atrial events. A transitional event of T-SR is detected when the mean period is within the range for sinus rhythm, but the rhythm is irregular. A similar criterion is used in detection of T-VT or T-AT.

Event 7 is detected when there is perfect synchrony, but the periods are too irregular or insufficient in number to be considered for other classes. Finally, event 8 is reserved for unclassified rhythms.

The event detection algorithm uses three features of synchrony, rate, and regularity of periods. The algorithm operates by "setting traps" for various events. It first tries to identify fibrillation or flutter (classes {5,6}). In absence of the two, it searches for fast beats (classes {3,4,7}), and then sinus rhythms (classes {1,2,7}). If none of the traps are triggered, the beat remains unclassified (class 8). Details of the detection algorithm are presented in Sheikhzadeh, Brennan, and So (2006).

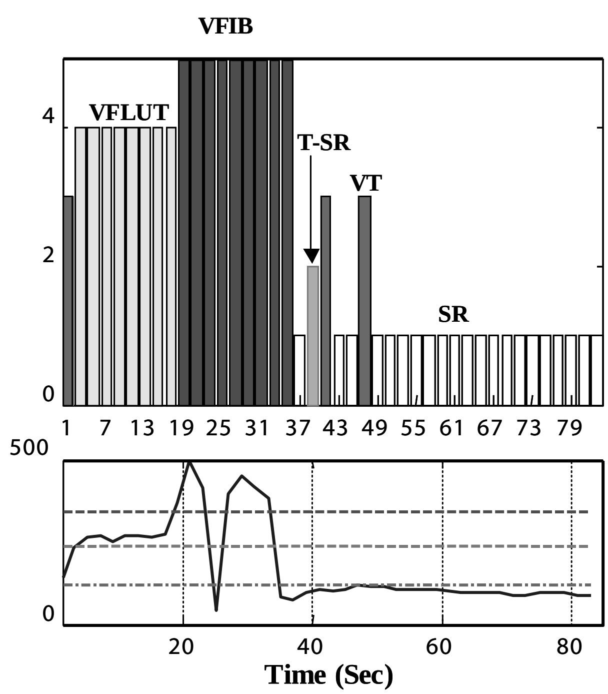
Performance Evaluation

All of the EGM data from Volume I of the Ann Arbor Electrogram Libraries (AAEL, 2006) were employed for system evaluation. Only the two bipolar EGM signals were utilized as they offer the best quality. The signals are recorded from Right Ventricular Apex and High Right Atrium, termed RVAb and HRAb in the AAEL documentation, respectively. The EGM signals were down-sampled from the original sampling frequency of *1000* Hz to *Fs=250* Hz, since our analysis and simulations revealed that it suffices. The detection system was optimized for a small (but diverse) set of input signals, and then tested with all available RVAb and HRAb EGM data, comparing the results to the AAEL annotation. The system parameters were tuned until there were no missed events, and the frame-by-frame classification was as accurate as possible for all types of rhythms. Figure 3 depicts a typical summary of the detection system output, including various detected events, and the beat-rate.

Detection in Noise

Practically, noise robustness is an important feature of an EGM processing system. To examine the effect of noise on system performance, noise was added to the EGM signal prior to detection. Employed noises included: 1-white noise, 2-lowpass, 3-bandpass, 4-highpass, and 5-tonal (60 Hz) noise. Noises 2-4 were filtered white noise with a 3-dB bandwidth of a quarter of full-band. Measurement of the signal-to-noise ratio

Figure 3. Algorithm performance for RVAb EGM signal (AAEL file: SET1, A177A26.SIG), top view: detected events in time, bottom view: detected rate in beats per minute (solid line) in time, and VFIB, VFLUT, and VT rate thresholds (dashed line, from top to bottom respectively)



(SNR) was based on tracking the short-term (four seconds) signal envelope (rather than power) of EGM and noise signals.

Event detection was performed at the noise levels reported in Table 1 for the whole EGM database, comparing the outputs to the ones for clean EGM signals. Events were clustered into two separate groups, {1-4} (SR/Tachy), and {5-6} (Fib/Flut). To further evaluate the performance, we compared the detection results in noise with the results in clean conditions, keeping in mind that it is especially vital to ensure that no event group is lost under noisy conditions. We measured the detection performance using frame counts of *TP, TN, FP,* and *FN* defined as:

1. *TP*, True Positive: Correct detection of {5-6}
2. *TP*, True Negative: Correctly not detecting {5,6}
3. *FP*, False Positive: Falsely detecting {5,6}
4. *FN*, False Negative: Falsely not detecting {5,6}

Then the Fib/Flut Positive Predictivity (+*P*) and Negative Predictivity (-*P*) were calculated as:

+P= TP/(TP+FP), -P= TN/(TN+FN).

The +*P* and –*P* measures for the five noise types are summarized in Table 1. As expected, white and lowpass noises have the most effect, and 60 Hz noise and high-pass noise have the least effect on performance. Overall, no block-event was missed or misrecognized with the five noise types in Table 1. The recognition errors in noise were sparse, and mostly confined to immediately before or after fibrillation or flutter events when the quality of EGM signal was already compromised.

Table 1. Positive and negative predictivities for noise types of 1-white, 2- lowpass, 3- bandpass, 4- highpass, and 5- 60 Hz at SNR levels of 15 dB for types 1-3 and 0 dB for 4-5

Noise Type	Positive Predictivity	Negative Predictivity
1	95.9	99.6
2	95.2	99.7
3	97.4	99.5
4	99.2	99.8
5	98.1	99.7

FUTURE TRENDS

Subband-based detection could be easily expanded to include morphology analysis and pattern-matching, based on subband features. This could outperform comparable time-domain methods, due to the superior and more robust representation of signal spectral features. Also, research is being conducted towards an ultra-low power implementation of the WOLA-based detection algorithm at power levels comparable to current ICDs (typically less than one μW). Initial results confirm the feasibility of this implementation. Subband-based detection methods may beneficially be combined with time-domain methods when time response is crucial (for example, in pacing).

CONCLUSION

As a result of subband decomposition and multitiered detection, subband-based EGM detection methods are superior to time-domain methods in terms of accuracy and robustness. Capturing synchrony of the subbands is key to the reliability and accuracy. The WOLA implementation, together with proposed simplified techniques, offers a feasible subband-based detection system for ultra-low resource platforms. Extensive testing of the proposed method with the AAEL database has shown excellent performance in terms of correct event detection and beat-rate measurement, even in fibrillation or flutter when the signal quality is compromised. Evaluation in noise and in the presence of far-field R-waves (not reported here) has also demonstrated significant robustness to noise and interference.

REFERENCES

AAEL. (2006), *Ann Arbor electrogram libraries*. Ann Arbor, MI. Retrieved February 1, 2008, from http://electrogram.com/

Afonso, V.X. et al. (1999). ECG beat detection using filter banks. *IEEE Transactions on Biomedical Engineering*, *46*(2), 192–202.

Astrom, M., Olmos, S., &Sornmo, L. (2006). Wavelet-based event detection in implantable cardiac rhythm management devices. *IEEE Transactions on Biomedical Engineering, 53*(3), 478–483.

Brennan, R. L., & Schneider, T. (1998). A flexible filterbank structure for extensive signal manipulations in digital hearing aids, In *Proceedings of IEEE International Symposium on Circuits and Systems* (Vol. 6, No. 1, pp. 569–572).

Brennan, R. L., & Schneider, T. (2003, August). *U.S. Patent No. 6,606,391.* Filterbank structure and method for filtering and separating an information signal into different bands, particularly for audio signals in hearing aids.

Brown, M.L., Christensen, J.L., & Gillberg, J.M. (2002). Improved discimination of VT from SVT in dual-chamber ICDs by combined analysis of dual-chamber intervals and ventricular electrogram morphology. *Computers in Cardiology, 29*(1),117–120.

Kohler, B.-U. et al. (2002). The principles of software QRS detection. *IEEE Magazine of Engineering in Medicine and Biology, 21*(1), 42–57.

Koyrakh, L.A., Gillberg, J.M., & Wood, N.M. (1999). Wavelet transform based algorithms for EGM morphology discrimination for implantable ICDs. *Computers in Cardiology, 26*(1), 343–346.

Rojo-Alvarez, J.L. et al. (2002). Discriminating between supraventricular and ventricular tachycardias from EGM onset analysis. *IEEE magazine of Engineering in Medicine and Biology, 21*(1),16–26.

Schaer, B. et al. (2000). Methods of minimizing inappropriate implantable cardioverter-defibrrilator shocks. *Current Cardiology Reports 2*, 346–352.

Sheikhzadeh, H., Brennan, R., & So, S. (2006). Cardiac rhythm detection and classification by WOLA filterbank analysis of EGM signals. In *Proceedings of IEEE Engineering in Medical and Biology Conference*.

KEY TERMS

Atrial Fibrillation (AFIB): A common heart rhythm disorder caused by a problem in the conduction of electrical impulses in the upper chambers, or atria, of the heart, causing it to beat chaotically.

Atrial Flutter (AFLUT): A heart rhythm disorder in which the upper chambers of the heart (the atria) contract faster than the lower chambers (the ventricles) in an organized, predictable pattern.

Atrial Tachycardia (AT) and Ventricular Tachycardia (VT): Refer to tachycardia originating from the ventricular and atrial chambers, respectively.

Cardioversion: The conversion of one cardiac rhythm or electrical pattern to another, almost always from an abnormal to a normal one. This conversion can be pharmacologic or electrical.

Defibrillation: The use of a carefully controlled electric shock, administered either through a device on the exterior of the chest wall or directly to the exposed heart muscle.

Electrogram (EGM): Sensed electrical activity of the heart by an electrode implanted in the heart.

Implantable Cardiac Defibrillator (ICD): A device that is put within the body, and is designed to recognize certain types of abnormal heart rhythms (arrhythmias), and correct them through cardioversion and defibrillation.

Supraventricular Tachycardia (SVT): AFIB and other rapid heartbeats that arise in the atria, or in the juncture between the atria and the lower chambers (ventricles).

Tachycardia: A rapid heart rate, usually defined as greater than 100 beats per minute.

Ventricular Fibrillation (VFIB): In VFIB, the heartbeat is rapid and chaotic, which causes the lower heart chambers, or ventricles, to go into a spasm.

U

Ultrasound Guided Noninvasive Measurement of Central Venous Pressure

Vikram Aggarwal
Johns Hopkins University, USA

Yoonju Cho
Johns Hopkins University, USA

Aniruddha Chatterjee
Johns Hopkins University, USA

Dickson Cheung
Johns Hopkins Bayview Medical Center, USA

INTRODUCTION

Central venous pressure (CVP) is a measure of the mean pressure within the thoracic *vena cava*, which is the largest vein in the body and responsible for returning blood from the systemic circulation to the heart. CVP is a major determinant of the filling pressure and cardiac preload, and like any fluid pump, the heart depends on an adequate preload to function effectively. Low venous return translates into a lower preload and a drop in overall cardiac output, a relationship described by the Frank-Starling Mechanism.

CVP is an important physiological parameter, the correct measure of which is a clinically relevant diagnostic tool for heart failure patients. In addition to other vitals such as heart rate and mean arterial pressure, accurate measures of central venous pressure through simple diagnostic instrumentation would provide physicians with a clear picture of cardiac functionality, and allow for more targeted treatment. Recent literature has also shown that measuring CVP can be an important hemodynamic indicator for the early identification and treatment of more widespread conditions, such as sepsis (Rivers, Nguyen, Havstad, & Ressler, 2001). With over five million patients (American Heart Association, http://www.americanheart.org/presenter.jhtml) in the U.S. presenting with heart failure-like symptoms annually, a current challenge for physicians is to obtain a quick and accurate measure of a patient's central venous pressure in a manner that poses minimum discomfort.

A novel noninvasive method to estimate CVP is to use ultrasound imaging in conjunction with a surface pressure measurement. Ultrasound is first used to visualize the *internal jugular* (IJ) vein below the skin on a patient's neck, and a custom pressure transducer is then used to detect the surface pressure required to collapse the IJ. The collapsing pressure is correlated to the central venous pressure and reported back to the operator. This proposed measurement procedure is suitable for emergency situations or primary care settings where rapid diagnosis of a patient's CVP is required, and mitigates the need for further invasive and costly procedures.

BACKGROUND

Central Venous Pressure is dependent on the ratio of venous blood volume (V_v) to venous compliance (C_v), as described in (1):

$$CVP = \frac{V_V}{C_V} \quad (1)$$

From this relationship, it is clear that an increase in blood volume within the thoracic *vena cava* will lead to greater pressure exerted on the vessel walls, and that a vessel with low compliance will experience greater pressure from an increase in blood volume than a vessel with high compliance. A variety of physiological conditions may influence CVP, due to changes in blood volume and/or compliance, and a few are presented here:

- **Hypervolemia:** Condition where overall blood volume is too high. Hypervolemia can arise from high salt intake or a failure in the kidney's ability to excrete salt and water from the body. Excess fluids seep into body tissue leading to edema (swelling), which can cause headaches and respiratory difficulties. The increase in blood volume increases central venous pressure.

- **Hypovolemia:** Condition where overall blood volume is too low. A common cause of hypovolemia is blood loss due to trauma, which leads to weakness and, in extreme cases, organ failure. The drop in venous blood volume decreases central venous pressure.
- **Heart Failure:** Condition where the heart is no longer able to maintain adequate blood circulation. Some causes of heart failure including coronary artery disease, valvular disease, and myocardial infarction. Low cardiac output leads to a backup of blood in the venous compartment, and increases central venous pressure.
- **Sympathetic Activation:** General increase in the activity of the sympathetic nervous system. Sympathetic activation, characterized as the "fight or flight" response, leads to increased heart rate, inhibition of the digestive system, and increased vascular tone. The latter decreases venous compliance, and increases central venous pressure.

Due to the high compliance of the thoracic *vena cava* and its inconvenient physiological location, direct measurement of CVP using a sphygmomanometer has long been considered impractical. Physicians have instead relied on direct measurements of intraluminal venous pressure, or crude noninvasive techniques for estimating CVP. This section explores both the traditional and more novel diagnostic approaches for measuring CVP, as a prelude to our discussion of the proposed measurement procedure and device.

Traditional Methods of Measuring Central Venous Pressure

a. Lewis (1930) describes one of the oldest methods to estimate CVP using jugular vein distension. When hypervolemic or cardiac failure symptoms manifest as increased venous blood volume, the extra blood pools and increases the size of the patient's jugular veins. The jugular vein's proximity to the surface of the neck makes it possible for a clinician to visually identify the swelling of neck veins, but this technique is not a true measurement of CVP since it is quite subjective. For one, the jugular veins are not always visible, especially when obscured by layers of fat, and venous distension is only readily apparent in cases of extremely high CVP. Therefore, this technique is useful for rapid diagnosis of a major increase in CVP, but not suitable for an objective measurement, rendering it largely inadequate for continuous monitoring.

b. Another noninvasive measurement technique that attempts to quantify CVP readings involves treating the jugular veins as manometer tubes to the right atrium, implying that higher pressure leads to a greater column of blood in the jugular veins (Lipton, 2000). The selected zero-reference point is the angle of Louis, located at the junction of the *manubrium* and the *sternum* 5 cm above the right atrium, because venous pressure at this point changes very little with posture. The height of the blood column can be determined by visually identifying jugular vein pulsations, which are oscillations in pressure caused by the cardiac cycle, and visible where the vein transitions from the distended (filled with blood) to the collapsed (no blood) state. The hydrostatic pressure that corresponds to the height difference between the point of fluctuations and the angle of Louis is used as a measure of CVP. In practice, this quantitative approach is often impractical, because the venous pulsations are only visible when the top of the blood column is in an area of the neck where the jugular passes close to the skin surface. If not, the clinician must raise the patient slowly upward from a supine position until the pulsations become visible, a process that is both physically taxing and time-consuming. Furthermore, if CVP is too high, then the top of the blood column may lie above the mandible, even with the patient sitting upright, making the fluctuations invisible. As with the previous method, layers of fat may also hinder visualization of the venous pulsations on the neck.

c. Numerous studies have shown that noninvasive methods fare poorly in quantifying CVP, because subjective measurements are difficult to standardize. Factors such as posture are hard to control in heart failure patients that must be lifted upright to observe jugular pulsations, which leads to highly variable and inaccurate measurements when compared to invasive recordings (McGee, 1998). Therefore, situations necessitating accurate CVP readings for prompt and correct treatment rely on invasive catheter-based measurements as the standard-of-care. Central venous access is the process of placing a venous catheter inside a major vein, which allows clinicians to obtain pressure readings from directly within the vessel lumen. A common insertion method is intravenous cannulation, which requires a larger insertion point in the vein to fit the cannula, and may lead to complications from excessive bleeding (Mitchell & Clark, 1979). A modified approach that uses a smaller insertion point is the Seldinger technique, which threads a small guide wire to the appropriate location, and then pushes the catheter along the wire (Seldinger, 1953). Although routinely performed, central venous accesses are susceptible to a variety of complications, such as difficulty locating the vein after the initial incision, overinsertion of wires and cannula leading to accidental puncture of an artery or lungs (Mitchell & Clark, 1979), and neurological complications from nerve lesions (Defalque & Fletcher, 1988). Due to the difficulty inherent in these techniques and the risk of complications, central venous access procedures require skilled medical personnel in a hospital setting.

Alternative Methods of Measuring Central Venous Pressure

a. Bloch, Krieger, and Sackner (1991) describe a method where an inductive neck plethysmograph is used to detect and record the small changes in neck dimensions that correspond to venous and arterial pulsations. In certain positions, such as the head-down position, the venous pulsation dominates in strength, whereas in the recumbent position, a mixed venous-arterial pulsation is detected. The device relies on measuring the height with respect to the phlebostatic axis, at which the waveform level changes from a venous to a venous-arterial or arterial configuration. Measurement of this height is a measure of CVP in cm H_2O. Their study indicates that this noninvasive method is accurate, compared to invasive catheter measurement. However, the measurement procedure is time-consuming and requires specialized equipment, like the plethysmograph, which is not as common as an ultrasound device.

b. Lipton (2000) describes a method where an ultrasound machine is used to locate the jugular venous pulsations. An ultrasonic probe is applied to the *internal jugular vein* (IJ), which lies very close to the neck surface, and provides a straight path to the right atrium. The IJ is a prime location for viewing jugular pulsations, and it is easily located by referencing the *sternocleidomastoid* muscle. This method allows the clinician to more easily determine the location of these pulsations, especially in situations where they are not visible on the surface of the neck. The vertical distance between this location and the right atrium is used as a measure of CVP.

c. Baumann, Marquis, Stoupis, Willenberg, Takala, and Jakob (2005) describe a method where an ultrasonic probe is custom-fitted with a quartz pressure-transducer. The probe is pressed against the IJ, until the vessel first starts to collapse on the ultrasound display, and this pressure is recorded as the bearing pressure. The probe is then pressed until the IJ fully collapses, and the final pressure is recorded. The difference between collapsing pressure and bearing pressure is the pressure needed to directly collapse the vessel. This quantity is equal to the pressure within the vessel lumen if the vein is assumed to be an ideal vessel following the Law of Laplace. While the assumption is imperfect, it is still a useful approximation compared to existing noninvasive approaches. The study noted high inter-operator variability and low precision, which may be improved by better design of the force transducer and standardized training.

A NOVEL NONINVASIVE APPROACH

A novel noninvasive method to estimate central venous pressure using ultrasound-guided surface pressure measurement is discussed here. Specifically, the device works in conjunction with an ultrasound machine and probe that is used to visualize the *internal jugular* (IJ) vein below the surface of the skin on a patient's neck. Once the vein is located, the device detects the pressure on the skin required to collapse the IJ, and correlates this value to a central venous pressure reading reported to the operator. This quick and noninvasive measurement is suitable for emergency situations or primary care settings where rapid diagnosis of a patient's CVP is required, and prevents the need for further invasive and costly procedures. The measurement procedure is also simple enough to be performed by operators without extensive medical training.

Measurement Principle

The measurement principle assumes that the IJ vein can be modeled as an ideal vessel according to the Law of Laplace as described in (2) and shown in Figure 1:

$$T = \frac{P \times R}{M} \qquad (2)$$

where T is wall tension, P is transmural pressure (difference between internal and external pressure), R is vessel radius, and M is wall thickness. When transmural pressure is zero, the tension falls to zero, and the vessel collapses. Conversely, if the vessel is directly collapsed, it can be inferred that the external pressure required to collapse the IJ is equal to the CVP. A custom handheld probe is responsible for measuring this applied external pressure, and is used alongside an ultrasound probe capable of visualizing the IJ. The *internal jugular vein* is located underneath the *sternocleidomastoid* muscle, and anteriolateral to the *common carotid artery*. The IJ is targeted because it provides a direct pipeline to the heart, and because of its anatomical proximity to the skin surface, making it easier to see on an ultrasound (Lipton, 2000).

After locating the IJ on the ultrasound monitor, the operator applies pressure with the handheld probe until the IJ first begins to collapse, and records this as the *initial pressure* required to collapse the superficial tissue between the skin and the vein. Since the tissue surrounding the IJ is not homogenous, the pressure applied at the skin surface will not directly collapse the IJ, but will be affected by factors like skin compliance and the presence of layers of obstructive tissue between the skin and vein. This step helps factor out the effects of surrounding soft tissue and varying initial skin compliance on the pressure measurement. The operator then continues to apply pressure until the IJ is completely collapsed, and records this final *collapsing pressure*. The central venous pressure is estimated as the difference between the final collapsing pressure and the initial pressure, as described in (3). In this manner, one can dynamically account for the surrounding soft tissue without need for calibration on each use.

$$CVP = P_{collapse} - P_{initial} \qquad (3)$$

This device utilizes the same principle of operation described in Baumann et al. (2005), but differs in its integration of a highly sensitive pressure transducer

Figure 1. Model of ideal vessel described by wall tension T, internal pressure P_i, external pressure P_o, and vessel radius R (Klabunde, 2004)

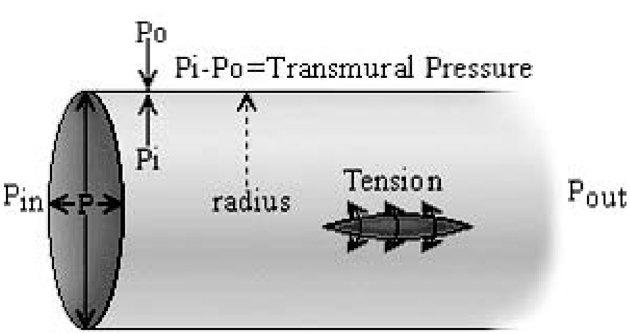
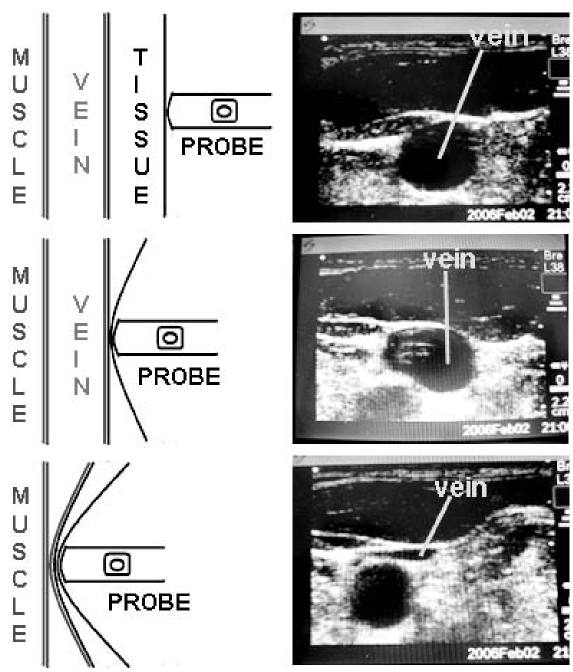


Figure 2. Left column shows progression of probe during each step of the measurement procedure; right column depicts actual ultrasound images visible to operator



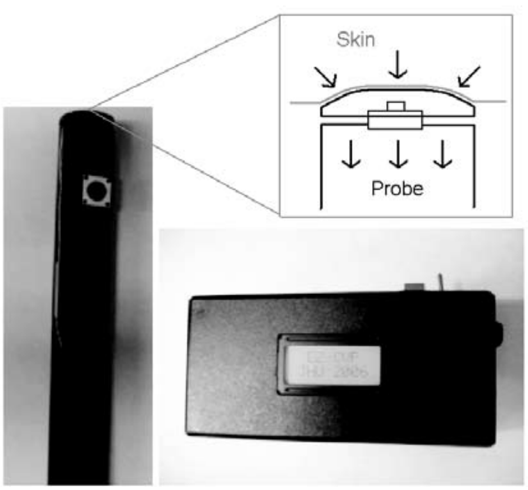
with a custom-designed probe that can be used in conjunction with any available ultrasound. Operation of the device is illustrated more clearly in Figure 2, with the probe and actual ultrasound images of the IJ during each stage of the measurement procedure.

Device Design

As shown in Figure 3, the device consists of a handheld probe with a pressure transducer and a control unit. The handheld probe is cylindrical in shape, and has a high-precision 5-lb load cell on the tip that detects the applied pressure and transmits this reading to the control unit. The probe also has a button along the longitudinal side of the probe body, which is used to initiate and end the CVP recordings. The probe tip (see inset of Figure 3) is designed to address how the collapsing force at the surface of the skin is transmitted to the IJ. Since the IJ is not exactly an ideal vessel, vessel compliance disperses the force somewhat. In order to minimize this, the probe tip is curved with a contact surface area of approximately 2 cm^2 that helps localize the applied force appropriately, and is an effective contact profile for compressing the IJ.

The control unit is battery-powered and consists of a microprocessor that analyzes the incoming load cell data, and an LCD that outputs the CVP readings.

Figure 3. Handheld probe (left) and control unit (right). Inset details design of custom probe tip, which takes distributed force from skin surface and channels it to the force transducer



The probe and control unit are meant to work in conjunction with an ultrasound machine and transducer for visualization of the *internal jugular vein*. The choice of ultrasound machine is up to the operator, since the described device is self-contained and does not require a particular ultrasound machine model. To aid in emergency diagnosis, a portable ultrasound is recommended; to aid in visualization accuracy, a high-frequency linear ultrasound transducer with low penetration depth is recommended.

FUTURE TRENDS

Ultrasound-guided therapies have gained significant momentum in the last decade, and are expected to continue to revolutionize health care technology. The U.S. ultrasound market is forecasted to grow at a compound annual growth rate (CAGR) of 5.9%, and reach over $1.89 billion by 2010 (Frost & Sullivan, 2004). This growth is driven primarily by the demand for portable units, as well as the increasing adoption of ultrasound technology among new end users such as cardiologists. Ultrasounds for cardiology, specifically, have a market with a CAGR of 7.6%, and will reach over $638 million by 2010—34% of the total market (Frost & Sullivan, 2004). And with manufacturers rapidly releasing smaller portable ultrasound machines that are easily transported and deployed in emergency settings, the future of medicine may see compact ultrasound machines replace stethoscopes as the diagnostic instruments of choice for physicians. The fast pace of research in ultrasound technology has brought faster and more accurate devices that are also equipped with features such as color flow Doppler mode, used commonly to detect blood flow velocity and direction in large vessels and cardiac chambers (Perry, Helmcke, Nanda, Byard, & Soto, 1987). Doppler ultrasound is one example of an imaging strategy that might support the proposed measurement procedure by providing operators with a less subjective means of determining complete collapse of the IJ during application of the probe. Considering the advances in ultrasound technology, our CVP measurement device fits well in this rapidly growing market, and may lead to widespread adoption of ultrasound devices by cardiologists and emergency room doctors, to primary care physicians and nurses.

The incorporation of a noninvasive measurement of central venous pressure into standard practice will enable clinical practitioners to use this vital parameter more often in their assessment of patient health. The benefits are clear in cases where patients present with heart problems and central venous access is a routine procedure. However, physicians have also begun using central venous pressure, along with other hemodynamic indicators, to help treat a wide variety of disorders. Hocking (2000) recommends that CVP should be measured in patients that are hypotensive, unresponsive to basic clinical management, or those requiring infusions of inotropes; Rivers (2001) used regular CVP measurements, among other indicators, as part of a goal-directed therapy to manage patients in septic shock; and Wang, Liang, Huang, and Yin (2006) report that patients who had their CVP lowered suffered less blood loss during liver resection. As CVP becomes an increasingly vital parameter for patient treatment, the advent of a reasonably accurate approach to noninvasively measure CVP can lead to significant improvements in standard of care in fields outside cardiology.

CONCLUSION

Clearly, there is a need for a system that provides a reliable noninvasive measure of a patient's central venous pressure. The real impact of the proposed measurement procedure and device will be felt in the 6,000 ICUs (NJHA, 2004) and 35,000 primary care facilities (HRSA, 2006) across the country, either as a necessary step in a life-saving procedure, or during measurement of key vitals during a standard exam. Critics who consider the price of an ultrasound too high for simply measuring CVP should also consider the numerous additional uses of ultrasound, including guidance in central venous access procedures (Denys, Uretsky, & Reddy, 1993; Troianos, Jobes, & Ellison, 1991) and visualization of organs such as the heart, liver, and kidneys. Another benefit of the proposed noninvasive approach is to mitigate the costs and risks associated with central venous access, a procedure that would be performed less often if a cheaper and safer alternative to measure CVP were available. The novelty of the technology may be a deterrent to its adoption by medical practitioners, but clinical studies of the technology and measurement procedure should be pursued to provide hard evidence of its utility. If the data is encouraging, industrial partners would be free to pursue product development and marketing in an open and highly profitable field. In conclusion, the direct ultrasound-guided noninvasive surface pressure measurement of CVP is a significant improvement over previous diagnostic techniques, and has the potential to make CVP measurements quicker, easier, and safer—thus more suitable for rapid diagnosis in emergency situations and primary care environments.

REFERENCES

Baumann, U., Marquis, C., Stoupis, C., Willenberg, T.A., Takala, J., & Jakob S.M. (2005). Estimation of central venous pressure by ultrasound. *Resuscitation, 64*, 193–199.

Bloch, K.E., Krieger, B.P., & Sackner, M.A. (1991). Noninvasive measurement of central venous pressure by neck inductive plethysmography. *Chest, 100*, 371–375.

Defalque, R.J., & Fletcher, M.V. (1988). Neurological complications of central venous cannulation. *Journal of Parenteral and Enteral Nutrition, 12*(4), 406–409.

Denys, B.G., Uretsky, B.F., & Reddy, P.S. (1993). Ultrasound-assisted cannulation of the internal jugular vein. A prospective comparison to the external landmark-guided technique. *American Heart Association, 87*, 1557–1562.

Frost & Sullivan. (2004). *U.S. ultrasound market – A675-50*. Palo Alto, CA.

Health Resources and Services Administration. (2006). *HRSA geospatial data warehouse – primary care service areas.* U.S. Department of Health and Human Services. Retrieved February 1, 2008, from http://datawarehouse.hrsa.gov/pcsa.htm

Hocking, G. (2000). Central venous access and monitoring. *Update on Anaesthesia, 12*(13), 1–6.

Klabunde, R. (2004). *Cardiovascular physiology concepts*. Lippincott Williams & Wilkins.

Lewis, T. (1930). Early signs of cardiac failure of the congestive type. *British Medical Journal, 1,* 849–852.

Lipton, B. (2000). Estimation of central venous pressure by ultrasound of internal jugular vein. *American Journal of Emergency Medicine, 18*(4), 432–434.

McGee, S.R. (1998). Physical examination of venous pressure: A critical review. *The American Heart Journal, 136*(1), 10–18.

Mitchell, S.E., & Clark, R.A. (1979). Complications of central venous catheterization. *Journal of Roentgenology, 133*(3), 467–476.

NJHA Quality Institute. (2004). Quest: News and updates from NJHA Quality Institute. *New Jersey Hospital Association, 3*(2).

Perry, G.J., Helmcke, F., Nanda, N.C., Byard, C., & Soto, B. (1987). Evaluation of aortic insufficiency by Doppler color flow mapping. *Journal of American Colleges of Cardiology*, 9952–9959.

Rivers, E., Nguyen, B., Havstad, S., & Ressler, J. (2001). Early goal-directed therapy in the treatment of severe sepsis and septic shock. *The New England Journal of Medicine, 345*(19), 1368–1377.

Seldinger, S.I. (1953). Catheter replacement of the needle in percutaneous arteriography: A new technique. *Acta Radiology, 39*(5), 368–76.

Troianos, C.A., Jobes, D.R., & Ellison, N. (1991). Ultrasound-guided cannulation of the internal jugular vein. A prospective, randomized study. *Anesthesia and Analgesia, 72*(6), 823–6.

Wang, W.D., Liang, L.J., Huang, X.Q., & Yin, X.Y. (2006). Low central venous pressure reduces blood loss in hepatectomy. *World Journal of Astroenterology, 12*(6), 935–9.

KEY TERMS

Cardiac Preload: Initial stretch of cardiac muscle before the systolic contraction phase; preload directly affects stroke volume and cardiac output through the Frank-Starling mechanism.

Cardiac Output: Total volume of blood ejected from left side of the heart into systemic circulation per minute.

Central Venous Pressure (CVP): Mean pressure inside the thoracic *vena cava*, near the right atrium of the heart; primary determinant of cardiac preload, and sensitive to changes in blood volume and venous compliance.

Compliance: Ratio of change in blood volume to change in intraluminal pressure; highly compliant vessels will respond to increased blood flow by distending, whereas low compliance vessels are more rigid, and will have increased pressure with more blood flow; compliance is regulated by the sympathetic nervous system, and affected by collagen build-up and short-term changes in vascular tone.

Frank-Starling Mechanism: Pioneering work of Otto Frank and Ernest Starling that provides a framework for relating cardiac muscle length, tension, and activation to parameters such as preload and cardiac output.

Heart Failure: Pathological condition whereby the heart has reduced cardiac output, leading to backup of blood volume in the venous system and increased central venous pressure; causes include coronary artery disease, valvular disease, and myocardial infarction.

Intravenous Cannulation: Insertion of a flexible tube or cannula into the venous compartment to allow for passage of a catheter; a procedure commonly performed when inserting a central intravenous line.

Plethysmograph: Instrument used to measure changes in volume of a vessel or organ.

Sphygmomanometer: An instrument for measuring arterial blood pressure, also known as a blood pressure cuff.

Ultrasound Imaging: Procedure by which high-frequency sound waves are reflected off internal tissues or organs to produce echoes, which are analyzed to form an image of body tissues, called a sonogram; ultrasound is commonly used in fetal monitoring, and more recently used in cardiology and radiology.

Using Object Oriented Technologies to Build Collaborative Applications in Healthcare and Medical Information Systems

A. Dwivedi
University of Hull, UK

R.K. Bali
BIOCORE, Coventry University, UK

Nilmini Wickramasinghe
Illinois Institute of Technology, USA

R.N.G. Naguib
BIOCORE, Coventry University, UK

INTRODUCTION

The adoption and diffusion of e-health and the application of IT in healthcare is increasing at a rapid rate. Both European and U.S. governments are making e-health a priority on their agendas. The technical infrastructure required to support initiatives such as community healthcare integrated networks (CHINs) and telemedicine efforts is often dependent upon connecting different types of computer networks, each running on different types of technologies so as to present to the user the image of a single virtual electronic health highway. It is generally agreed that current software development technology cannot deliver this due to limitations of restricted scalability, fragmented management, and inflexibility in providing business support.

One of the potential solutions may be the use of Object Oriented (OO) technology. This article explores the feasibility of combining OO technologies with healthcare based workflow management systems (WFMS). We introduce the concept of workflow technologies and discuss the main advantages and limitations of WFMS. We detail the circumstances in which the use of WFMS could be considered and the technological factors necessary for its successful implementation.

We also present an Object Management Group (OMG) model, analysing it in the context of the support offered for WFMS. The main advantages and disadvantages of the model are discussed. A workflow management coalition (WFMC) model is then contrasted with the OMG management model in order to identify the architectural differences between them. We focus on the relationship between workflow concepts and the position of the two reference models (WFMC and OMG) and on the use of UML in the design of information systems. We conclude by summarizing our findings on the extent to which OO technology can be used to build collaborative applications in healthcare and medical information systems.

BACKGROUND

During his State of the Union Address in January 2004, President George Bush affirmed the intention of the government to emphasize the role of technology in administration and delivery of healthcare in the United States (Bush, 2004). Similar sentiments have been voiced by the European leaders (Global Medical Forum Foundation, 2005; The Oslo Declaration on Health, 2003), and the World health organization ("E-Health in Eastern Mediterranean," 2005; A Health Telematics Policy, 1997). Both European and U.S. authorities define their initiatives primarily in terms of medical information technology centering on computerized patient record (CPR) or, in more acceptable parlance, the electronic health record (EHR). See Brailer and Terasawa (2003).

WHO's platform statement (A Health Telematics Policy, 1997) speaks of "health telematics policy," an all inclusive term that incorporates not only EHR but essentially all healthcare services provided at a distance and based on the use IT.

While implementation of these concepts is preeminently realistic in the context of EU and the U.S.A., the WHO plan appears, for many reasons, a combination of a list of good ideas and delineation of significant obstacles that make the good ideas seem almost futuristic. In response to the inefficiency of the highly fragmented programs to address even the most urgent aspects of healthcare across the globe, a demand for the development of a new rule set (Banjeri, 2004; Barnett, 2004; Olutimayin, 2002; Onen, 2004) governing the future actions began to emerge—the quest for the "doctrine of global health."

To address this void, von Lubitz and Wickramasinghe developed the doctrine of "networkcentric healthcare" (von Lubitz & Wickramasinghe, 2006a, 2006b, 2006c), which calls for the development of interconnected information grids that, together, constitute a powerful and well-structured network that facilitates information sharing among all participants within the operational continuum (Cebrowski & Garstka, 1998; Stein, 1998). Consequent to improved information sharing is the enhancement of its quality and integrity which, in turn, escalates the level of situational awareness that is the foundation for efficient, real-time collaboration among the involved entities, their self-synchronization, and operational sustainability which leads to a dramatic increase in mission effectiveness (Cebrowski & Garstka, 1998).

As described by von Lubitz and Wickramasinghe (2006a, 2006b, 2006c), networkcentric healthcare operations must be conducted within the intersecting territory of three mutually interconnected and functionally related domains (Garstka, 2000):

- The *physical domain* which encompasses the structure of the entire environment healthcare operations intend to influence directly or indirectly, for example, elimination of disease, fiscal operations, political environment, patient and personnel education, and so forth.
- The *information domain* which contains all elements required for generation, storage, manipulation, dissemination/sharing of information, and its transformation and dissemination/sharing as knowledge in all its forms. It is here that all aspects of command and control are communicated and all sensory inputs gathered.
- The *cognitive domain* relates to all human factors that affect operations, such as education, training, experience, political inclinations, personal engagement (motivation), "open-mindedness," or even intuition of individuals involved in the relevant activities. Difficulties in metrics relevant to the cognitive domain notwithstanding, a body of experimental studies begins to emerge that will, ultimately, provide close quantitative relationships to other domains that govern healthcare operations space (Abel-Smith, 1989; Back & Oppenheim, 2001; Bodner et al., 1986; Newby, 2001; Roberts & Clifton, 1992; Wetherell et al., 2002).

The essential and enabling technology element of NCHO is the Worldwide Healthcare Information Grid (WHIG) that allows full and unhindered sharing of information among individual domains, their constituents, and among constituents across the domains (von Lubitz & Wickramasinghe, 2006a, 2006b, 2006c). In order to perform such a function, the WHIG must consist of an interconnected matrix of ICT systems and capabilities (including communication platforms, data collection, storage, manipulation/dissemination, and sharing), associated processes (such as information and knowledge storage and retrieval, management and their dissemination/sharing), people (e.g., healthcare providers/investigators, administrators, economists, politicians, lawyers, ICT personnel), and agencies (governmental and Non-Governmental Organizations or NGSs) at local/national/international level.

However, von Lubitz and Wickramasinghe have not detailed the technological make up of the WHIG which is the backbone of NCHO. We contend that the ultimate use of Object Oriented (OO) technologies as we discuss in this article will be to provide the technological backbone to such initiatives as NCHO or smaller scale initiatives such as community healthcare integrated networks (CHINs) or e-health and telemedicine applications. To fully appreciate the power and benefits that OO technologies bring to effecting superior healthcare delivery, it is necessary first to understand the evolution of IT applications in healthcare and the key challenges to date.

EVOLUTION OF IT APPLICATIONS IN HEALTHCARE

In the 1970s, the focus of IT applications in healthcare was to facilitate better healthcare administration, particularly in routine administrative tasks such as the calculation of patient charges for reimbursement . In the 1980s, the focus shifted to the development of clinical systems to aid in patient diagnosis and treatment (Johns, 1997; Rao, 2001). This trend continued until the late 1990s. A distinguishing feature of this trend was that most of the IT applications in healthcare were developed for very specific uses such as standalone software applications (Rao, 2001). This lack of interoperability has today become a significant challenge for the healthcare sector.

Simultaneously, in the last 40 years, there has been a significant evolution in system architectures that form the substrate of IT applications. Organizations have evolved from a centralised to a distributed computing architecture (Ganti & Brayman, 1995). This change of emphasis in system architectures has been accompanied by the emergence of client/server and object oriented (OO) technologies. By the 1990s, these two technologies had signalled the coming of age of the distributed computing architecture (Ganti & Brayman, 1995).

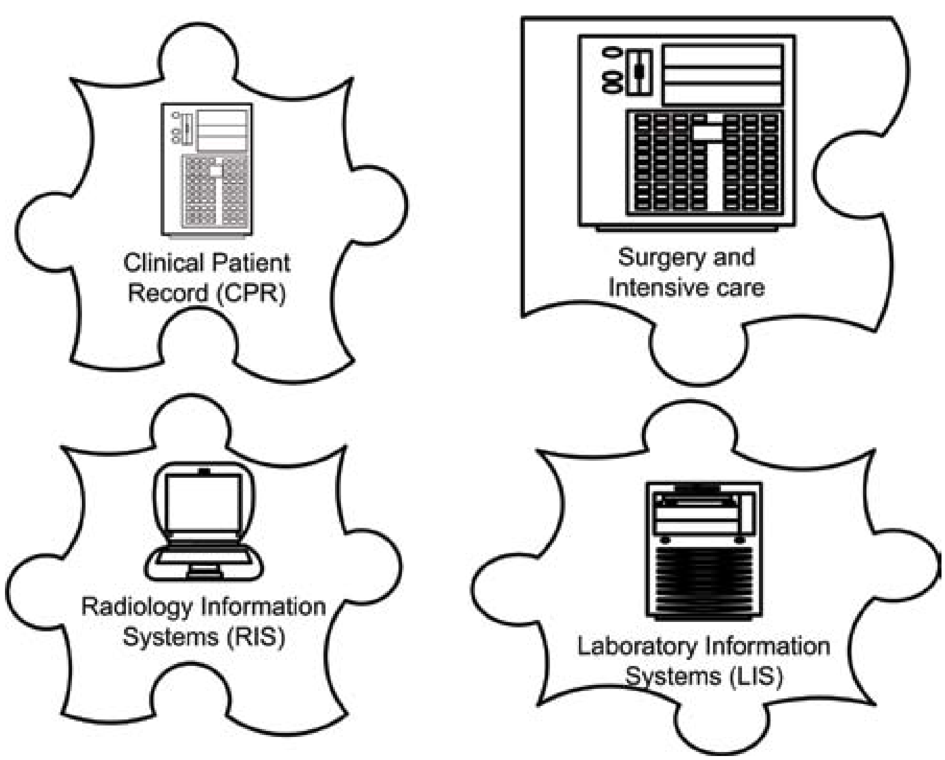
Modern IT applications have evolved to provide basic support for services to aid knowledge processing, problem-solving, and coordination by incorporating concepts such as intelligent design systems, workflow management, and intelligent agents (Ganti & Brayman, 1995). These changes (i.e., evolution from a centralised to a distributed computing architecture technologies) have had implications on the healthcare sector. Even IT applications in healthcare have undergone a drastic change (Ohe, 1998; Pollard & Hammond, 1998; Rao, 2001).

In the centralised computing architecture, IT applications in healthcare were characterised by being written in a proprietary language, specifically made to function on a particular hardware platform, and, most importantly, had very limited interfaces (connections, ability to communicate) with other healthcare information systems (clinical, diagnostic, etc.) (see Figure 1). Moreover, they did not have to ability to interconnect with common desktop applications (word processing, etc.).

This lack of interoperability between different but related healthcare systems (see Figure 1) has been cited as one of the main challenges facing the healthcare sector (Ferrara, 1998; Greenberg & Welcker, 1998; Harrington, 1993). This problem has been further aggravated by the advent of modern IT applications in healthcare that are centred on the power of the Internet (Egger, 2000; Weber, 1999).

The use of OO technologies in conjunction with protocols such as the HyperText Transport Protocol (HTTP) and the Wireless Application Protocol (WAP), with devices such as Personal Digital Assistant (PDA), have allowed patients and doctors to remain in close contact

Figure 1. Limited interfaces between IT healthcare apps



(Dwivedi, Bali, James, & Naguib, 2001; Farkas, 1999; Pollard & Hammond, 1998). These new technologies (OO, HTTP, and WAP) have collectively triggered the connectivity revolution in healthcare. Simultaneously, these new technologies have been accompanied by a revolution in medical electronics and human genome research. It is estimated that in the next 10 years, more then 100,000 medical devices will be launched in the healthcare sector (Egger, 2000).

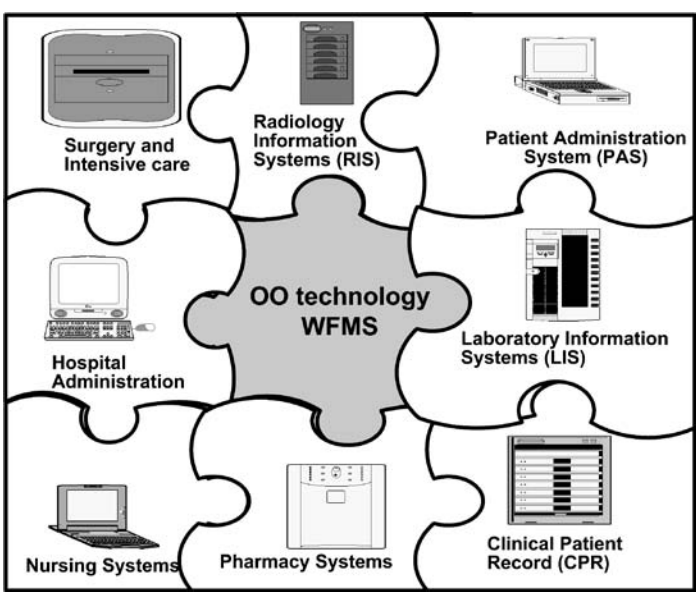
Advances in the application of computer technology have largely driven the biomedicine revolution in the healthcare industry. As a result of the above revolution, modern day healthcare industry is witnessing "the emergence and proliferation of entire new scientific disciplines—molecular biology, economics, bioinformatics and proteomics—which are revealing the secrets of genes, proteins and their functions" (Watanabe, 1999, p. 27). This has implications for the healthcare process.

Advances in pharmaceutical developments will in turn be driven largely by these new scientific disciplines. Our understanding of the human genome could result in the discovery of new treatments based on new pharmaceutical compounds. It has been estimated that in the near future new pharmaceutical compounds could replace 50% of today's in-patient services (Egger, 2000). It is clear that the healthcare system of the new millennium will be computing based and would revolve around Information Technology (Egger, 2000).

The benchmark for survival for healthcare institutions in the future will shift from the current emphasis on "facilities, range of services,or clinicians associated with the organization" to the "experience that the healthcare institution has in providing that care" (Egger, 2000, p. 22), which in turn would require an effective integration of all the different systems associated with healthcare treatment, diagnosis, and delivery. The vision of the healthcare system of the new millennium is one where all the different information systems relating to healthcare delivery are connected seamlessly and appear to the user as a virtual system, see Figure 2 (Johns, 1997; Mercer, 2001).

Object Oriented Technology (OOT) and Workflow Management Systems (WFMS) are the two most promising technologies that stand out from the entire set of all the technologies that can provide a seamless integration for a virtual healthcare system for the future (see Figure 2). We contend that these two technologies, in combination, hold the potential of transforming the vision of the healthcare system of the new millennium into a reality. As mentioned earlier, one of the biggest paradigm shifts in computing was the evolution of system architectures from centralised computing architecture to a distributed computing architecture. This paradigm shift brought out to the forefront the enabling power of client/server computing (Khanna, 1994).

Figure 2. Vision of an integrated healthcare system in the new millennium



EVOLUTION OF OOT

One of the main reasons for the popularity of the client/server revolution is the sharing of data, applications, and the processing power between different clients and the server, thereby improving the efficiency of the system. OOT, and in particular distributed objects, are likely to be the driving force that will transform the vision of an integrated distributed computing architecture into reality; see Table 1 (Orfali, Harkey, & Edwards, 1996).

OOT is a software development methodology that uses a commonly accepted body of knowledge that includes notation diagrams to represent the design of a software application. One of the main advantages of adopting the OO methodology is that it reduces the software development cycle; it offers support in extending the application further by allowing reuse of existent software components.

In addition to the above, OO methodology offers more flexibility in building and interfacing the proposed /existing application with other applications. This not only makes it easier to maintain software applications produced under the OO methodology, but also substantially reduces the cost of software development (Athwall & Moreton, 1995).

Prior to the mid 1980s, all the major types of Hospital Information Systems (HIS) were based on the centralized computing architecture (Khanna, 1994; Ohe, 1998). These legacy HIS used to consist of a number of powerful mainframes which were linked to a very large number of dumb terminals (Ohe, 1998).

One of the main problems in the 1990s was how to utilise these legacy systems with the modern HIS. This gave impetus to the need for having a computing architecture that enabled coordination between different resources (hardware and software) that was available on a network, and the concept of the distributed computing architecture that could enable the above to become a reality (Khanna, 1994).

Table 1. Evolution of client/server with OOT

	1982	1986	1990	1995	1998 +
First Wave	File Servers				
Second Wave		Database Servers			
			Groupware		
Third Wave				Distributed Objects	

DISTRIBUTED COMPUTING BASED CLIENT/SERVER ARCHITECTURE

One of the main disadvantages of early client/server technologies was that there was a clear separation between the two, resulting in the creation of two monoliths instead of one (Orfali et al., 1996). Any vision of a healthcare system in the new millennium would feature knowledge-intensive applications. These applications would evolve from a convergence of digital and multimedia technologies and all of which are likely to run a large number of heterogeneous, autonomous, and distributed information systems (Tsiknakis, Chronaki, Kapidakis, Nikolaoul, & Orphanoudakis, 1997).

This poses severe limitations for existing client/server applications, most of which are based on old technologies (see Table 1). Distributed objects are being regarded as a technology that holds a lot of potential in empowering client/server applications to deal with the challenge of repeatedly integrating a large amount of multimedia clinical data from a wide number of heterogeneous, autonomous, and distributed HIS within a very limited time span (Orfali et al., 1996; Tsiknakis et al., 1997).

Distributed objects through the use of components can achieve the above by subdividing today's monolithic client/server applications into self-managing components that can interact with each other and move across networks and operating systems (Orfali et al., 1996). Distributed objects take full advantage of having the data and business logic which is encapsulated within them to interact with other (old) legacy applications via object wrappers (Orfali et al., 1996). This makes the system more flexible as it takes full advantage of the concept of modularity in software development.

DEFINITION AND INTERRELATIONSHIP BETWEEN DISTRIBUTED OBJECTS AND COMPONENTS

Distributed objects can be defined as an independent piece of created code that is visible to other remote clients irrespective of the language and compiler used to create it. It has the ability to be accessed by other remote clients via method invocation. To put it in simple terms, a distributed object is a blob of intelligence that can live anywhere on the network and which can message each other transparently (Orfali et al., 1996).

There is no precise definition of the term component as it represents diverse features depending upon the perspective taken. However, any definition of components must include the fact that it is a reusable piece of software that is independent of any application (Orfali et al., 1996). Components can be a black box (i.e., they support polymorphism and encapsulation but not inheritance) or white box that supports all three mentioned concepts of traditional objects (Orfali et al., 1996).

A broad overview of the relationship between components and objects could be verbalised to state that each component is an object that is not bound to a particular program, computer language, or implementation (Orfali et al., 1996). The interrelationship between distributed objects and components is that all distributed objects are components but not all components are objects or distributed objects, for example, OLE and black box components (Orfali et al., 1996). This is because components have synergistically evolved from three paradigms (1) the distributed computing architecture, (2) the client/server revolution, and (3) the OO revolution.

Components are different from traditional objects in the sense that components are standalone objects that can plug-and-play across networks, applications, languages, tools, and operating systems (Orfali et al., 1996). As each component has its own intelligence and data, the use of components would allow developers and users to create applications on the fly very easily. Moreover, by using components based on distributed objects on a client/server platform, it is possible to very simply build a large number of customised software applications. The use of components represents the ultimate form of client/server distribution and in the near future almost all desktops would be using client/server distribution (Orfali et al., 1996).

Table 2. Evolution of DBMS

Search capabilities/ multi-user support.	Relational DBMS	Object Relational DBMS
	File Systems	Object Oriented DBMS
	Data complexity/ extensibility	

If we look at the evolution of database management systems (DBMS) in the context of their suitability for multi-user support whilst handling complex data types, the following picture emerges (see Table 2) (Connolly & Begg, 1998).

A previous study has noted that almost all modern day HIS are based on traditional relational database systems. However, it has been acknowledged that relational database systems cannot deal with the large volumes of clinical data that come in different formats (audio, video, images, and multimedia) (Farkas, 1999). This study examined both extended relational and object relational databases (see Table 2) and on their ability to store and access clinical data that is multimedia in nature, and come to the conclusion that in a HIS context, object-oriented databases is the best solution (Farkas, 1999). All the existing HIS use client/server technology with the server running a traditional relational database that cannot store and supply multimedia data (X-ray images, etc.) effectively, while their client side often uses peculiar user interfaces that are difficult to govern and are usually not platform independent (Farkas, 1999). The main rationale behind a relational database application is that it establishes modularity between data—modularity which is not established not because it is an essential requisite, but only because it makes the system easier to implement and manage (Farkas, 1999).

WFMS

A Workflow Management System (WFMS) is an ideal software application that assists in the processing of tasks. At a basic level, workflow can be thought of as a series of actions that help in the automation of tasks.

Since workflow refers to a series of tasks, it implicitly refers to scheduling or routing (i.e., which is the best possible way of performing a particular group of tasks). Since it refers to tasks which have to be performed, it would require responsibility at a basic level. Thus a workflow model can be thought to be a value adding process which is carried out in an efficient and effective manner with clear responsibilities for each of the participants in the process (Workflow Management Coalition, 1995). The WFMS uses process instances to instantiate each business process, and each process instantiation represents a workflow. This gives the WFMS the facility of having several workflows of the identical business process, at the same time. This means that each of them (each process instantiation) can act in a different manner. The advantages here are comparable to the advantages given by multitasking and object orientation (Workflow Management Coalition, 1999).

The Workflow Management Coalition (WFMC) (1995) have provided a Workflow Reference Model (WRM) that has the following five main interfaces:

1. **Process definition tools:** This interface provides specifications for process definition data and interchange.
2. **Other workflow enhancement services:** This interface provides support for interoperability between different workflow systems.
3. **Invoked applications:** This interface provides support for interaction with a variety of IT application types.
4. **Workflow client applications:** This interface provides support for interaction with the user via desktop functions.
5. **Administration and specifications for process definition and data interchange:** This interface provides support for system monitoring and metric functions to facilitate the management of composite workflow application environments.

One of the main weaknesses of the WRM model recommended by the WFMC is that it does not support workflow implementation "across a heterogeneous, distributed infrastructure" (Santanu, Edwin, & Jarir, 2001).

This drawback forces the workflow system architecture to be a client/server system architecture where all the main workflow services are run from the workflow server and, as such, it becomes "monolithic" and "centralised." The main disadvantage of the centralised workflow system architecture is that it "does not address the needs of distributed workflows on a WAN" and thereby prevents the workflow to function in a distributed manner (Santanu et al., 2001). This is a very serious drawback, particularly considering that the future of workflows is connected with the Web or, at minimum, across multiple sites.

The Object Management Group (OMG) is an international body whose objective is "to promote the theory and practice of OOT" by establishing industry standards (Athwall & Moreton, 1995). OMG has challenged the WRM by providing its own version of workflow systems.

The OMG architecture uses CORBA as middleware (a software that helps in assuring client/server communications) that enables distributed objects to cooperate. The OMG architecture uses Object Request Brokers (ORBs) to provide the basic functionality for establishing communication between objects.

This is achieved via two interfaces: (1) the application specific interface and/or (2) the Domain interface. All the CORBA facilities given by the OMG can be classified into: (1) Object services (OSs), which are families of system services, that are universal and have standardised interfaces, and (2) Common Facilities (CFs), which specify the specific services for a particular application domain (e.g., healthcare) (Pollard & Hammond, 1998). In the OMG model, workflow is a part of its CORBA facilities and falls under common facilities (Leymann & Roller, 2000).

The main difference between these two models is that, under the WFMC model, the workflow engine has a centralist character, while in the OMG architecture workflows are just objects. The WFMC model uses five interfaces whilst the OMG model uses just objects, thus fully exploiting the CORBA facilities (both object services and common facilities).

FUTURE TRENDS

Modern day healthcare organizations have realised that in the future their survival would depend upon their ability to give the caregiver access to such information that would enable the caregiver to deliver personalised clinical diagnosis and treatment in real-time in very specific clinical contexts (a process termed as Information Therapy). Information therapy is defined as the "prescription of specific, evidence-based medical information to a specific patient, caregiver, or consumer just in time to help someone make a specific health decision or behaviour change" (Kemper & Mettler, 2002, p. 17). This vision has been translated into concepts like Integrated Delivery System (IDS) and Community

Health Information Networks (CHIN) (Lang, 1997; Mercer, 2001; Morrissey, 2000).

IDS refers to a HIS that is a business model based on computing technologies such as OO "to share key data, with partners and providers, that will allow faster and more accurate decision making… to deliver care to a broader population with fewer requirements for expensive and scarce resources" (Lang, 1997, p. 18). CHINs are integrated HIS based upon a combination of different technology platforms that are connected to enable support for data sharing amongst different healthcare providers (Mercer, 2001).

Both IDS and CHIN are very similar in nature and both refer to an integrated network for allowing the delivery of personalised healthcare. We argue that in the near future component technology would be the driving force behind all HIS as it would support new ways of combing islands of knowledge, and present the knowledge acquired to the user as an integrated whole. In addition to the above, component technology would also provide support for the use of modern computer techniques (such as Intelligent Data Mining tools, WFMS) to discovering previously undiscovered patterns on a case-to-case basis, thereby bringing to the fore a truly integrated healthcare system that supports delivery of personalised healthcare.

On a grander, global scale, such component technology and the integral role for the OO-platform we have discussed is essential in supporting networkcentric healthcare operations (NCHO) as outlined by von Lubitz and Wickramasinghe (2006a, 2006b, 2006c). The constructing of NCHO is clearly a large task that requires to coordination of several players at a global level. However, the WHIG (as discussed earlier), the world healthcare information grid and backbone of NCHO, cannot be constructed without utilising these component technologies and OO platform. To date actualising the structure of the WHIG has not been discussed.

We contend that the ultimate use of Object Oriented technologies for healthcare as we have discussed in this article will be to provide the technological backbone to such initiatives as NCHO. NCHO represents the new paradigm for healthcare delivery while the OO technologies coupled with the WFMC model is an essential enabling criterion and critical success factor fore realising superior healthcare delivery.

CONCLUSION

Old legacy HIS were predominantly used to allow computerization of information to aid in providing financial information and to assist the healthcare managers in having centralized control over different healthcare activities (Lang, 1997).

Healthcare institutions in the 1990s have realized that the old model does not work as the nature of information required is different. Contemporary healthcare institutions are under ever-increasing pressure to find new ways of reducing healthcare costs whilst simultaneously increasing administrative and clinical efficiency to provide a superior level of quality relating to patient care (Carlos & Comaford, 1998). Modern day HIS have a number of clinical information systems each of which is committed to diverse healthcare providers (nurses, doctors, etc.) in different clinical disciplines (radiology, surgery, etc.) to effectively ensure coordination among these necessitates the integration of hospital-wide available information (Tsiknakis et al., 1997).

There is an increasing realization that an integrated HIS is an infrastructure prerequisite if healthcare institutions are to successfully meet the above mentioned challenge (Farkas, 1999; Ohe, 1998; Pollard & Hammond, 1998; Tsiknakis et al., 1997). The old model of HIS cannot meet the above challenge as the old legacy HIS running monolithic computing applications based upon centralized computing architecture cannot ensure an integrated HIS (Lang, 1997; Tsiknakis et al., 1997).

The distributed computing architecture is the only means of providing the integration between different heterogeneous systems that are autonomous and distributed (Lang, 1997; Pollard & Hammond, 1998; Tsiknakis et al., 1997). Another factor which holds a lot of potential in ensuring that the vision of an integrated HIS becomes a reality is the Unified Modelling Language (UML). Developed by the OMG, UML is a graphical and object-oriented notation methodology for describing processes in a form that helps both developers and users (Botelho, 2000a, 2000b). The advent and acceptance of UML as the defacto industry standard has made possible a standard way to depict graphically the design of an information system that can be understood by all the stakeholders at any stage of the software development cycle (Botelho, 2000b). Despite being of OMG specification, UML is independent of any middleware technology, that is, OMG's-CORBA, Microsoft's Component Object Model (COM), and Sun's Enterprises JavaBeans (EJB) (Sutherland, 1998). It is clear that, in the new millennium, an innovative approach would have to be adopted to ensure that the vision of a virtual integrated HIS becomes a reality. We conclude by reiterating that the current gap between different islands of heterogeneous HIS can be bridged if the technologies discussed above (OO, Component Technology, Distributed Computing, WFMS, and UML) are concurrently adopted.

REFERENCES

Abel-Smith, B. (1989). Health economies in developing countries. *J. Trop. Med. Hyg.*, *92*, 229–241.

Athwall, A.K., & Moreton, R. (1995). Use of object-oriented technology in healthcare. In B. Richard (Ed.), *Current perspective in healthcare computing* (Vol. 2, pp. 583-584).

Back, J., & Oppenheim, C. (2001). A model of cognitive load for IR: Implications for user relevance feedback interaction. *Information Res.*, *2*. Retrieved from http://InformationR.net/ir/6-2/ws2.html

Banjeri, D. (2004). The people and health service development in India: A brief overview. *Int. J. Health Serv. 34,* 123-42.

Barnett. T.P.M. (2004). *The Pentagon's new map*. New York: G.P. Putnam (pp. 1-435).

Bodner, G.M., & McMillen, T.L.B. (1986). Cognitive restructuring as an early stage in problem solving. *J. Res. Sci. Teaching*, *23*, 727–737.

Botelho, J. (2000a). The latest look in systems management. *Telephony, 238*(23), 252-258.

Botelho, J. (2000b). The object-ive viewpoint. *Telephony, 239*(19), 74-80.

Brailer D.J., & Terasawa, A.B. (2003). *Use and adoption of computer-based patient records.* California Healthcare Foundation (pp. 1-42).

Bush, G. (2004). State of the Union Address, The White House. Retrieved February 4, 2008, from http://www.whitehouse.gov/news/release/2004/01/20040120-7.html

Carlos, L.G., & Comaford, K.M. (1998). *Future trends in healthcare enterprise-wide knowledge distribution.* Paper presented at the Annual Conference of Healthcare Information and Management Systems Society (pp. 231-242).

Cebrowski, A.K, & Garstka, J.J (1998). Network-centric warfare: Its origin and future. *U.S. Nav. Inst. Proc. 1,* 28-35.

Connolly, T.M., & Begg, C.E. (1998). *Database systems: A practical approach to design, implementation, and management.* Harlow: Addison-Wesley.

Dwivedi, A., Bali, R.K., James, A.E., & Naguib, R.N.G. (2001). *Workflow management systems: The healthcare technology of the future?* Paper presented at the 23rd Annual International Conference of the IEEE - Engineering in Medicine and Biology Society (EMBS), Istanbul, Turkey (pp. 3887-3890).

Egger, E. (2000). Market memo: How technology is changing the healthcare system. *Healthcare Strategic Management, 18*(1), 1, 20-23.

E-health in the eastern Mediterranean, WHO-EMRO. (2005). Retrieved February 4, 2008, from http://www.emro.who.int/his/ehealth/ehealthPlan.htm

Farkas, Z. (1999). Internet technologies in hospital information systems. *Internet Research, 9*(2), 3.

Ferrara, F.M. (1998). The CEN healthcare information systems architecture standard and the DHE middleware. A practical support to the integration and evolution of healthcare systems. *International Journal of Medical Informatics*, *48*(1-3), 173-182.

Ganti, N., & Brayman, W. (1995). *The transition of legacy systems to a distributed architecture.* New York: Wiley.

Garstka, J.J. (2000). Network centric warfare: an overview of emerging theory. *Phalanx*, *4*, 28–33.

Global Medical Forum Foundation (2005, April 4-6). *GMF IV: New frontiers in healthcare.* Retrieved February 4, 2008, from http://www.globalmedicalforum.org/gmf_iv/session3.asp

Greenberg, D.S., & Welcker, B. (1998). Active X based standards for healthcare integration. *International Journal of Medical Informatics, 48*(1-3), 183-190.

Harrington, J. (1993). *Recommendations for an object oriented healthcare information model.* Paper presented at the Progress in Standardization in Healthcare Informatics (pp. 52-59).

A health telematics policy in support of WHO's health-for-all strategy for global health development: Report of the WHO group consultation on health telematics. (1997, December 11-16). World Health Organization, Geneva, 1998.

Johns, P.M. (1997). Integrating information systems and healthcare. *Logistics Information Management, 10*(4), 140-145.

Kemper, D.W., & Mettler, M. (2002). Information therapy: A tale. *Health Forum Journal, 45*(1), 16-20.

Khanna, R. (1994). *Distributed computing: Implementation and management strategies*. Englewood Cliffs, NJ: Prentice Hall.

Lang, R.D. (1997). The promise of object technology: What's in it for healthcare? *Journal of Healthcare Resource Management, 15*(10), 18-21.

Leymann, F., & Roller, D. (2000). *Production workflow: Concepts and techniques*: Prentice-Hall International.

Mercer, K. (2001). Examining the impact of health information networks on health system integration in Canada. *Leadership in Health Services, 14*(3), 1-30.

Morrissey, J. (2000). The evolution of a CHIN. *Modern Healthcare, 42*.

Newby, G.B. (2001). Cognitive space and information space. *J. Am. Soc. Info. Sci. Technol, 12*, 1026–1048.

Ohe, K. (1998). *A hospital information system based on common object request broker architecture (CORBA) for exchanging distributed medical objects - an approach to future environment of sharing healthcare information.* Paper presented at the MEDINFO 98, Amsterdam.

Olutimayin J. (2002). Communication in healthcare delivery in developing countries: Which way out? *Pac. Health Dialog. 9*, 237-241.

Onen, C.L. (2004). Medicine in resource-poor settings: Time for a paradigm shift? *Clin. Med. 4,* 355-360

Orfali, R., Harkey, D., & Edwards, J. (1996). *The essential distributed objects survival guide.* New York: Wiley.

The Oslo declaration on health, dignity and human rights. (2003). 7th Conference of European Health Ministers, Oslo, Norway, June 12-13, 2003. Retrieved February 4, 2008, from http://odin.dep.no/hod/norsk/aktuelt/taler/monister/042071-990209/dpk-bn.html

Pollard, D.L., & Hammond, W.E. (1998). *Object technology: Raising the standards for healthcare information systems.* Paper presented at the MEDINFO 98, Amsterdam (pp. 217-221).

Rao, S.S. (2001). Integrated healthcare and telemedicine. *Work Study, 50*(6), 222-228.

Roberts, L.W., & Clifton, R.A. (1992). Measuring the cognitive domain of the quality of student life: An instrument for faculties of education. *Can. J. Educ., 2*, 176–191.

Santanu, P., Edwin, P., & Jarir, C. (2001). *Essential requirements for a workflow standard*. Retrieved February 4, 2008, from http://www.jeffsutherland.org/oopsla97/santanu.html

Stein P. (1998). *Observations on the emergence of network centric warfare*. Retrieved February 4, 2008, from http://www.dodccrp.org/research/ncw/stein/_observations/steincw.htm

Sutherland, J. (1998). Why I love the OMG: Emergence of a business object component architecture. *ACM StandardView, 6*(1), 4-13.

Tsiknakis, M., Chronaki, C.E., Kapidakis, S., Nikolaoul, C., & Orphanoudakis, S.C. (1997). An integrated architecture for the provision of health telematic services based on digital library technologies [Special issue on digital libraries in medicine]. *International Journal on Digital Libraries, 1*(3), 257-277.

von Lubitz, D., & Wickramasinghe, N. (2006a). Healthcare and technology: The doctrine of networkcentric healthcare. *Int. J. Electronic Healthcare (IJEH), 4,* 322-344.

von Lubitz, D., & Wickramasinghe, N. (2006b). Networkcentric healthcare: Outline of the entry portal concept. *Intl J. of Electronic Business, 1*(1), 16-28.

von Lubitz, D., & Wickramasinghe, N. (2006c). Networkcentric Healthcare and bioinformatics. *Intl J. Expert Systems, 30,* 11-23.

Watanabe, A.M. (1999). Living with innovation: Issues for medicine in the 21st century. *Executive Speeches, 14*(3), 26-31.

Weber, D.O. (1999). Web sites of tomorrow: How the Internet will transform healthcare. *Health Forum Journal, 42*(3), 40-45.

Wetherell, J.L., Reynolds, C.A., Gatz, M., & Pedersen, N.L. (2002). Anxiety, cognitive performance, and cognitive decline in normal aging. *J. Gerontol (B) Psych. Sci. and Soc. Sci., 57*, 246–255.

Workflow Management Coalition. (1995). *The workflow reference model*. Retrieved February 4, 2008, from http://www.wfmc.org/standards/docs/tc003v11.pdf

Workflow Management Coalition. (1999). *The workflow management coalition terminology & glossary* (Document No. WFMC-TC-1011). Retrieved February 4, 2008, from http://www.wfmc.org/standards/docs/TC-1011_term_glossary_v3.pdf

KEY TERMS

CHINs: Community healthcare integrated networks.

E-Health: The use of Internet-based technologies to facilitate healthcare delivery.

Object Oriented (OO) Technology: Technologies that support/utilise object oriented programming.

Telemedicine: The aiding of medical support from a distance via technology.

Workflow Management System (WFMS): Systems that support effective and efficient workflow.

VEMH: Virtual Euro-Mediterranean Hospital for Evidence-Based Medicine in the Euro-Mediterranean Region

G. Graschew
SRU OP 2000, Max-Delbrueck-Center and Robert-Roessle-Klinik, Charité - University Medicine Berlin, Germany

T.A. Roelofs
SRU OP 2000, Max-Delbrueck-Center and Robert-Roessle-Klinik, Charité - University Medicine Berlin, Germany

S. Rakowsky
SRU OP 2000, Max-Delbrueck-Center and Robert-Roessle-Klinik, Charité - University Medicine Berlin, Germany

P.M. Schlag
SRU OP 2000, Max-Delbrueck-Center and Robert-Roessle-Klinik, Charité - University Medicine Berlin, Germany

INTRODUCTION

Telemedicine aims at equal access to medical expertise irrespective of the geographical location of the person in need. New developments in Information and Communication Technologies (ICT) have enabled the transmission of medical images in sufficiently high quality, which allows for a reliable diagnosis to be determined by the expert at the receiving site (Pande, Patel, Powers, D'Ancona & Karamanoukian, 2003; Lacroix et al., 2002).

At the same time, however, these innovative developments in ICT over the last decade bear the risk of creating and amplifying a digital divide in the world, creating a disparity between the northern and southern Euro-Mediterranean areas (Dario et al., 2005; Graschew, Rakowsky, Roelofs & Schlag, 2003a; Graschew, Roelofs, Rakowsky & Schlag, 2004).

In recent years, various institutions have started Euro-Mediterranean telemedicine projects (EMISPHER, www.emispher.org/; BURNET, www.burnet.org/; PARADIGMA, www.paradigmamed.org; EMPHIS, www.emphis.org/; EUMEDGEN, www.eurogene.org; ODISEAME, www.odiseame.org; EUMEDCONNECT, www.eumedconnect.net/; GALENOS, www.rrk-berlin.de/op2000/Deutsch/projekte/galenos.html), which were intended to foster a cooperation between the European EU-Member countries and the Mediterranean countries (Graschew, Roelofs, Rakowsky, & Schlag, 2002; Graschew, Roelofs, Rakowsky, & Schlag, 2003b; Rheuban & Sullivan, 2005; Wootton, Jebamani, & Dow, 2005).

All these projects have demonstrated how the digital divide is only part of a more complex problem: the need for integration. Therefore, provision of the same advanced technologies to the European, Mediterranean, and Adhering countries should be the final goal for contributing to their better dialogue for integration.

BACKGROUND

Already in the framework of the EMISPHER project (Euro-Mediterranean Internet-Satellite Platform for Health, Medical Education and Research, EUMEDIS pilot project 110, see www.emispher.org, 9/2002-12/2004, cofinanced by the EU in the framework of the EUMEDIS program), an Internet-satellite platform for telemedicine in the Euro-Mediterranean region was established and put into operation (Graschew, Roelofs, Rakowsky & Schlag, 2005). Other telemedicine systems are used; for example, for tele-ultrasound in rural areas where telementoring by live videoconferencing allowed guiding the ultrasound technician to record additional images of the patient (O`Neill, Allen & Brockway, 2000); for clinical assessment of pediatric burns, which showed a good agreement between the face-to-face consultation and seeing the patient via videoconference (Smith, Kimble, Mill, Bailey, O'Rourke & Wootton, 2004); and for home telecare services likely to improve quality of health services

(Guillen et al., 2002). Other systems are described in Sable (2002); Latifi, Peck, Porter, Poropatich, Geare & Nassi (2004); and Eadie, Seifalian & Davidson (2003). Currently the network consists of 10 partners in Morocco, Algeria, Tunisia, Egypt, Turkey, Italy, Greece, Cyprus, France, and Germany (see Figure 1) and offers applications in the domains of medical e-Learning (courses for undergraduates, graduates, young medical doctors, etc., in real time and asynchronously), real-time telemedicine (second opinion, demonstration and dissemination of new techniques, telementoring, etc.), and e-Health (medical assistance for tourists and expatriates). The EMISPHER network serves as a basis for the introduction of a virtual hospital in the Euro-Mediterranean region.

Due to the experience in the exploitation of previous European telemedicine projects and, in particular, to activities carried out in the framework of the EUMEDIS program, an open Euro-Mediterranean consortium would like to propose the Virtual Euro-Mediterranean Hospital (VEMH) initiative.

VEMH aims to facilitate and accelerate the interconnection and interoperability of the various services being developed (by various organizations at various sites) through real integration. This integration must take into account the social, human, and cultural dimensions, and strive toward common approaches but open and respectful of cultural differences: multilateral cooperation instead of aid.

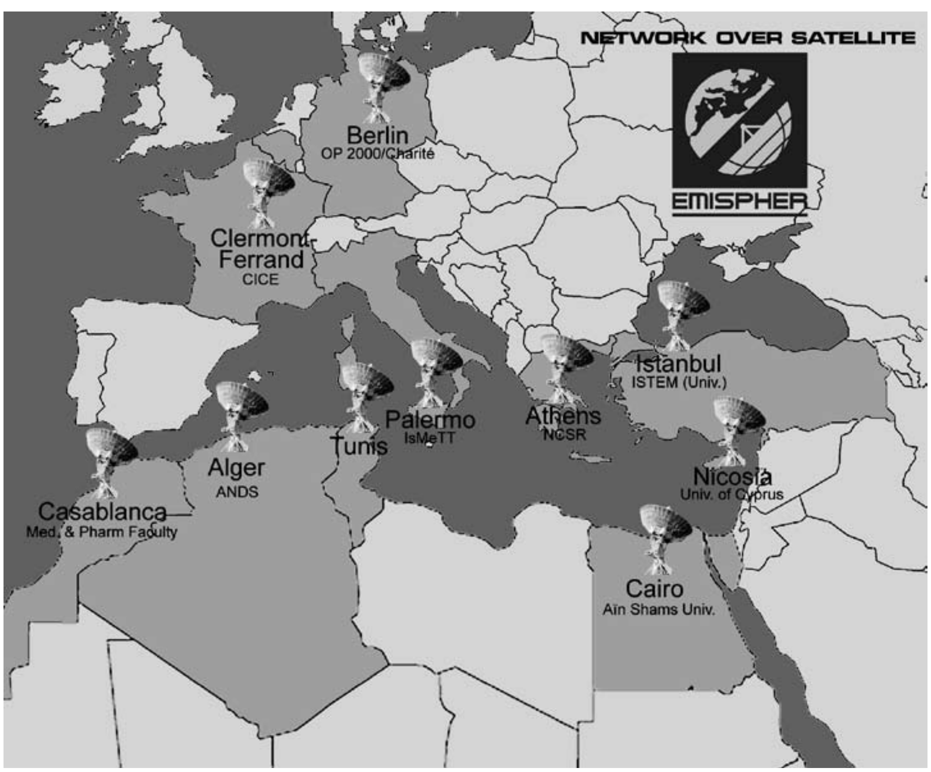
THE VIRTUAL EURO-MEDITERRANEAN HOSPITAL (VEMH)

VEMH will provide a heterogeneous integrated platform consisting of satellite links and terrestrial links for the application of various telemedical services. Evidence-based medicine will be integrated more and more in the following three main dedicated medical services.

E-Learning

In the project, the Mediterranean Medical University (MeMU) will be developed. The leading medical centers

Figure 1. Centers of excellence in the EMISPHER project interconnected by a satellite-based network for bridging the digital divide in the Euro-Mediterranean health care area



integrated in the network provide pedagogical material and modules for synchronous and asynchronous e-learning in their medical specialties: endoscopic surgery, gynecology-obstetrics, infectious diseases, liver and multi-organ transplantation, and so forth. The exchange among the partners of various countries allows for improved qualification of undergraduate and graduate students, hospital staff, general practitioners, and so forth (see Figure 2).

Real-Time Telemedicine

VEMH offers second opinions, tele-teaching and tele-training, tele-mentoring, and optimization of the learning curve. These real-time interactive telemedical applications contribute to improved quality of patient care and to accelerated qualification of medical doctors in their respective specialty. Thus, this international network of distributed but integrated competence contributes directly and indirectly to improved health care (see Figures 3 and 4).

Medical Assistance

As tourism constitutes a substantial economical factor in the Mediterranean region, and because of the increasing mobility of the population, continuity of care through improved medical assistance is of major importance for improved health care in the Euro-Mediterranean region. The introduction of standardized procedures, integration of the platform with various local and national communication systems, and training of medical staff involved in medical assistance allow for shared management of files related to medical assistance (medical images, diagnosis, workflow, etc.).

Evidence-Based Medicine

The next service in VEMH will be Evidence-Based Medicine (EBM). EBM cannot substitute the agreement on concrete and individual therapeutic procedures to be followed; however, even for the individual patient, EBM provides a more solid base for individual decisions. EBM contributes to better disease management.

Figure 2. Interactive tele-education session between IsMeTT (Palermo) and UCY (Cyprus) in the EMISPHER Project

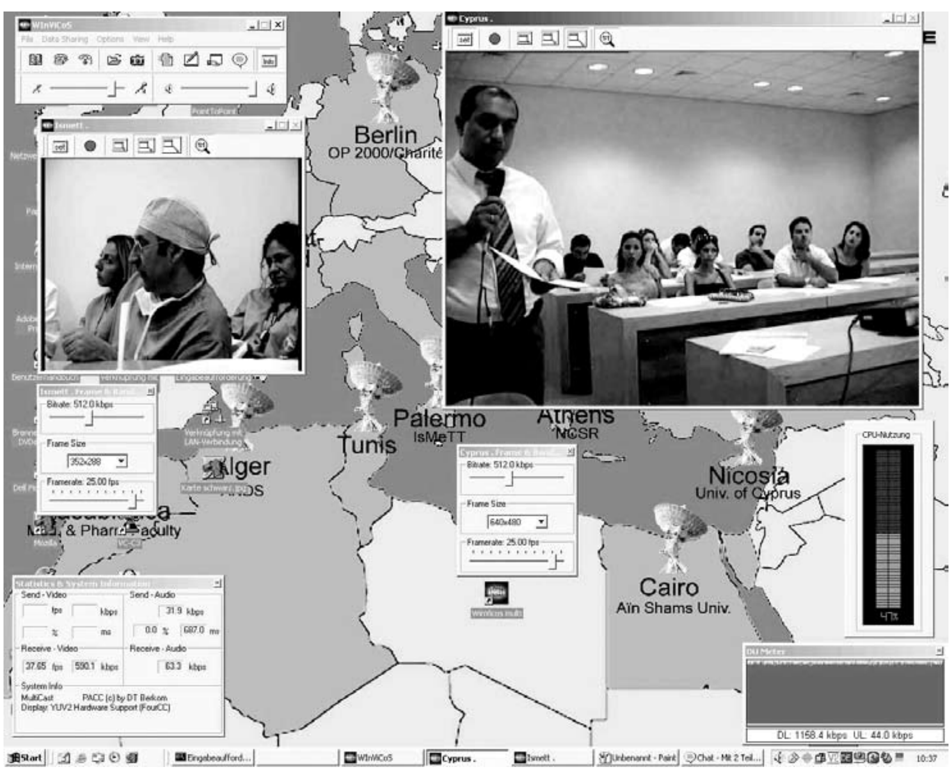


Figure 3. Interactive multipoint tele-consultation during laparoscopy (OP 2000 (Berlin), FMPC (Casablanca), and CICE (Clermont-Ferrand) in the EMISPHER Project

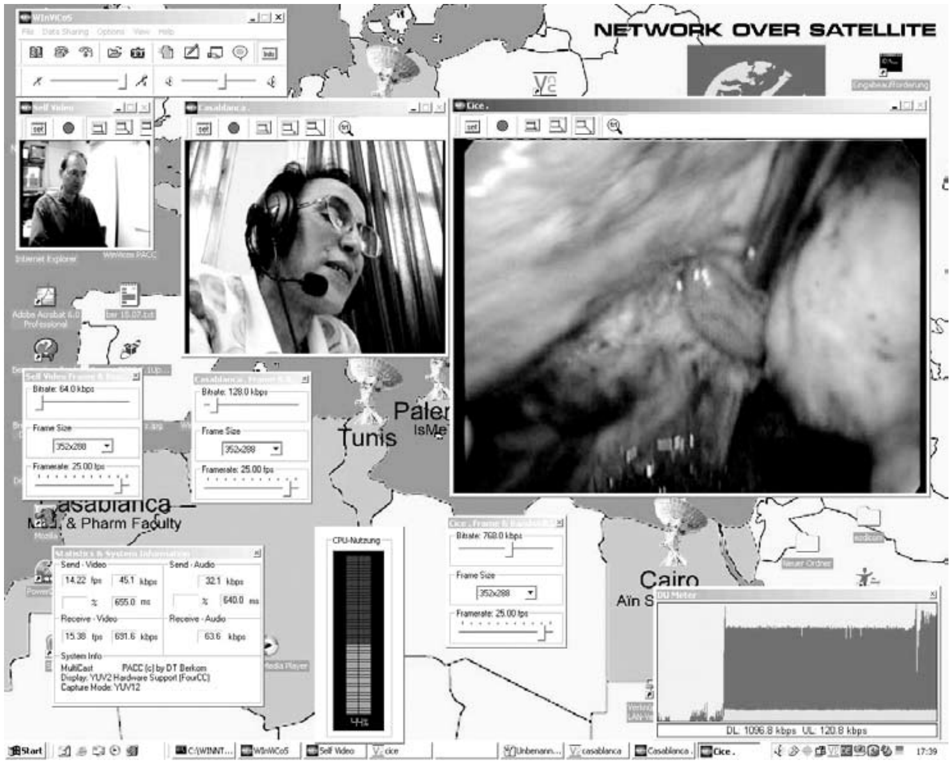
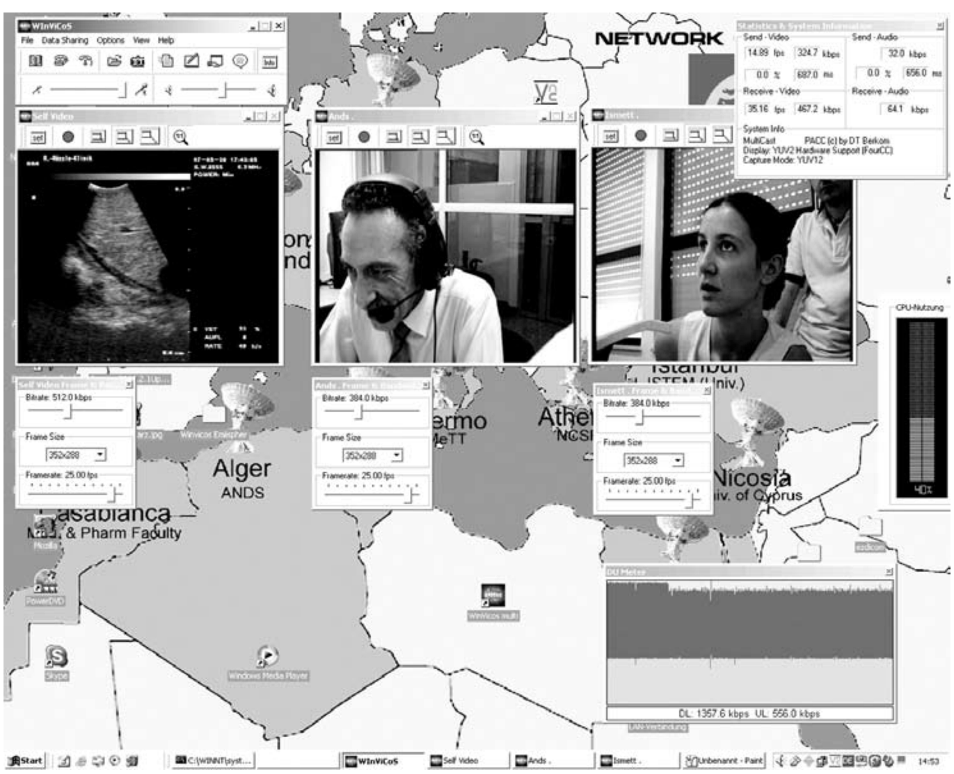


Figure 4. Teleconsultation between OP 2000 (Berlin), ANDS (Algiers), and IsMeTT (Palermo); a live video sequence of an ultrasound investigation is transmitted and diagnosed by the connected experts



Health care policymakers can also expect great potential from EBM.

Fellowship Program

VEMH will offer individual grants to young medical doctors for training in clinical and scientific organizations of the network. The scope of the fellowship program is to allow young medical doctors to develop and gain experience in a multicultural and multidisciplinary environment.

Methodology and System Architecture of VEMH

Improvement of the methodology of the VEMH network of competence is based on the management of clinical outcomes as well as the integration of various telemedical solutions in one platform to support many medical applications. The individual requirements of the user can be represented as matrix columns and the various applications with integrated technological solutions as matrix rows, while low access costs serve as selection criteria.

A modular network for the integration and optimization of various applications is created. The newly developed methodologies should be usable for more than one disease. The orientation from technique-oriented toward medical-need-oriented solutions serves as optimization criteria.

To approach the identified telemedicine services, the following requirements need to be considered:

- **Traffic:** Intensive applications such as teleconsultation, telediagnosis, and teletraining require real-time interactivity of the audio and video stream
- **High quality of images and video transmission:** Min. 386 kbps up to 1 Mbps
- **Guaranteed bandwidth:** Transmission of medical data does not allow for transfer delay and quality loss
- **Guaranteed confidentiality of patient data.**

Therefore, satellites, also considering the non-fully operating terrestrial infrastructures of the relevant sites and the bandwidth requirements, are the most fitting infrastructure for the envisaged set of VEMH services.

Given the use of already installed terminals and the already defined procedures for the network adaptation and configuration, it will be possible to start the VEMH activities at the beginning of the project.

Evaluation of the VEMH

The justification of the various VEMH services and applications will be assessed by using a comprehensive evaluation methodology. In particular, this will examine various outcomes, including clinical, organizational, and economic outcomes. The criteria under which such services can be evaluated are based upon the work of Bashshur (1995) and displayed in Table 1. The rationale for using this type of methodology is to ensure that the services or applications are capable of having an

Table 1. Criteria for the evaluation of VEMH services/applications

	Criteria							
	Evaluation of Service or Application							
Variables	**Safety**	**Reliability**	**Acceptability**	**Design**	**Effectiveness**	**Efficiency**	**Cost**	**Impact**
Clinical	✓	✓			✓	✓		✓
Economic					✓	✓	✓	✓
Business			✓		✓	✓	✓	
Technical/Standards	✓	✓	✓	✓				✓
Organizational	✓	✓	✓			✓		✓
Social/Political			✓		✓			
Medicolegal/Ethical	✓		✓					

immediate and positive impact upon patient care in the VEMH region.

Each service or application will be properly evaluated using the appropriate quantitative/qualitative tools and measurement processes. The quantitative assessment tools that will be used are cost-benefit analysis and cost-effectiveness analysis. Cost-benefit analysis compares the discounted future streams of incremental program benefits with incremental service or application costs; the difference between these two streams is the net social benefit of the service or application. The aim of the analysis is to identify whether a service or application's benefits exceed its costs. Cost-effectiveness analysis is designed to find the minimum cost of meeting a given target, and, unlike cost-benefit analyses, no attempt is made to estimate the benefits.

Both tools are required for a rational appraisal of the case for adoption, or ongoing use, of various services or applications. Failure to conduct these analyses can result in a waste of human and financial resources that ultimately have an effect on patient care.

Risks, Barriers, and Facilitators

Although the market is promising and the technology is ready (or nearly ready) to be used, the take-up and commercialization of telemedicine services are still uncertain due to a number of barriers:

External:

- Decision-making in health is fragmented with respect to procurement policies, which hinders progress.
- Establishing and building confidence with physicians takes time.
- It still seems that it will be one to two years more before full deployment and therefore a critical mass can be built.
- How the services are packaged and attitudinal changes from within the health sector are also necessary.
- There are numerous organizational issues.

Economic:

- It is difficult to predict the break-even level of services and therefore to determine the minimum level of investment required for break-even services.
- There is a lack of investment in health, and therefore, significant investment is still required.
- Considerable investment in improved management, training and education of personnel, redesigning of care, and logistic processes is necessary.

Legal:

- Ethical issues
- Regulatory framework issues
- Reimbursement issues
- Lack of standardization

Technical:

- Software development is still evolving.
- IP levels of connectivity are also required as a minimum.
- All the services need to be fully integrated into one platform.

In various pilot projects, the technology has been put into place, the necessary applications have been developed, and they proved to be used successfully and meet the needs of its end users. Yet numerous trials and demonstrations carried out during the projects have also highlighted a certain number of issues that potentially could hinder the commercialization process of e-services in the medical sector.

In fact, a market analysis report, "The Emerging European Health Telematics Industry" (February 2000, prepared by Deloitte & Touche for the European Commission Directorate General for the Information Society), also insists that the telemedicine market growth will be dependent upon a number of vital conditions and enabling factors already noted. The report also proposes that four key action lines should be initiated in parallel: consolidation on the supply side, technical integration, investment on the demand side, and accompanying measures that could be the enabling factors required to allow the health telematics market to achieve substantial and exponential growth.

The solutions to such barriers to commercialization do not lie within the scope of the technology and appear to be generic to those up and running telehealth activities worldwide. It is only in North America that new legislation has been introduced to respond to the particular needs of this practice; this legislation is not yet being applied on a federal level, but rather on a state-by-state basis. However, the organizational and cultural aspects that must accompany any new form of practice need the input of other actors both on the governmental, legal, and political scale. Telehealth is not a simple extension of current health systems and cannot be perceived as such.

Solutions to the barriers can be consolidated under the following three main categories:

- Awareness of telemedicine and telehealth as a medical practice
- The need for common standards and norms
- The need for specific legislation

FUTURE TRENDS

The application of advanced telecommunications networks and services to health care practices is still evolving and maturing, and hence, telemedicine has yet to become an established part of day-to-day health care practice in many regions and countries. The main barriers to the development of the telehealth market relate to the organization of health care services and, consequently, support to the connection to and usage of advanced telecommunications networks and services by health care establishments, which is primarily a matter for the public health authorities to consider.

However, the promise of telemedicine to provide equal access to medical expertise irrespective of the geographical location can only be met when not merely the patient's data are transferred, but also a telepresence is created bringing patient and remote experts together using ICT. Besides general interactivity between the two sites, features like telehaptic, telesensation, and remote control of medical devices (e.g., telerobotics) are prerequisites for a real telepresence.

Due to the distributed character of VEMH, data, computing resources as well as the need for these are distributed over many sites in the Virtual Hospital. Therefore, Grid infrastructures and services become inevitable for successful deployment of services such as acquisition and processing of medical images (3D patient models), data storage, archiving and retrieval, and data mining (especially for evidence-based medicine). In order to achieve this, conventional Grid technology has to be expanded to cover not only local computing resources but to a dimension of organization-spanning integrated networks.

CONCLUSION

VEMH will foster cross-Mediterranean cooperation among the leading medical centers of the participating countries by establishing a permanent medical and scientific link. Through the deployment and operation of an integrated satellite and terrestrial interactive communication platform, VEMH will provide medical professionals in the whole Euro-Mediterranean area access to the required quality of medical services, depending on the individual needs of each of the partners.

The applications in the areas of e-Learning, real-time telemedicine, and improved medical assistance contribute to an improved level of health care in the whole Euro-Mediterranean region and build the basis for the introduction of Evidence-Based Medicine.

REFERENCES

Bashshur, R.L. (1995). On the definition and evaluation of telemedicine. *Telemedicine Journal, 1*, 19–30.

Dario, C., et al. (2005). Opportunities and challenges of eHealth and telemedicine via satellite. *Eur J Med Res, 10* (Suppl. I). Proceedings of ESRIN-Symposium, Frascati, Italy, 1–52.

Eadie, L.H., Seifalian, A.M., & Davidson, B.R. (2003). Telemedicine in surgery. *Br J Surg, 90*, 647–658.

Graschew, G., Roelofs, T.A., Rakowsky, S., & Schlag P.M. (2002). *Broadband networks for interactive telemedical applications*. Proceedings of the APOC 2002, Applications of Broadband Optical and Wireless Networks, Shanghai. Proceedings of SPIE, Vol. 4912, 1–6.

Graschew, G., Rakowsky, S., Roelofs, T.A., & Schlag, P.M. (2003a). *Telemedicine as a bridge to avoid the digital divide world*. Proceedings of the Fortbildungsveranstaltung und Arbeitstagung Telemed 2003, Berlin, 7–8.

Graschew, G., Roelofs, T.A., Rakowsky, S., & Schlag, P.M. (2003b). *Telepresence over satellite*. Proceedings of the 17th International Congress on Computer Assisted Radiology and Surgery, London. *International Congress Series, 1256*, 273–278.

Graschew, G., Roelofs, T.A., Rakowsky, S., & Schlag, P.M. (2004). Interactive telemedicine as a tool to avoid a digital divide of the world. In L. Bos (Ed.), *Medical care and compunetics* (pp. 150–156). Amsterdam: IOS Press.

Graschew, G., Roelofs, T.A., Rakowsky, S., & Schlag, P.M. (2005). Überbrückung der digitalen teilung in der Euro-Mediterranen gesundheitsversorgung—das EMISPHER-Projekt. In A. Jäckel (Ed.), *Telemedizinführer Deutschland* (pp. 231–236). Minerva, Darmstadt: Ausgabe.

Guillen, S., et al. (2002). User satisfaction with home telecare based on broadband communication. *J Telemed Telecare, 8*, 81–90.

Lacroix, A., et al. (2002). International concerted action on collaboration in telemedicine: Recommendations of the G-8 Global Healthcare Applications Subproject-4. *Telemed J E-Health, 8*, 149–157.

Latifi, R., Peck, K., Porter, J.M., Poropatich, R., Geare III, T., & Nassi, R.B. (2004). Telepresence and telemedicine in trauma and emergency care management. *Stud Health Technol Inform, 104*, 193–199.

O'Neill, S.K., Allen, D., & Brockway, P.D. (2000). The design and implementation of an off-the-shelf, standards-based tele-ultrasound system. *J Telemed Telecare, 6* (Suppl. 2), 52–53.

Pande, R.U., Patel, Y., Powers, C.J., D'Ancona, G., & Karamanoukian, H.L. (2003). The telecommunication revolution in the medical field: Present applications and future perspective. *Curr Surg, 60*, 636–640.

Rheuban, K.S., & Sullivan, E. (2005). The University of Virginia telemedicine program: Traversing barriers beyond geography. *J Long-Term Eff Med Implants, 15*, 49–56.

Sable, C. (2002). Digital echocardiography and telemedicine applications in pediatric cardiology. *Pediatr-Cardiol, 23*, 358–369.

Smith, A.C., Kimble, R., Mill, J., Bailey, D., O'Rourke, P., & Wootton, R. (2004). Diagnostic accuracy of and patient satisfaction with telemedicine for the follow-up of paediatric burns patients. *J Telemed Telecare, 10*, 193–198.

Wootton, R., Jebamani, L.S., & Dow, S.A. (2005). E-health and the Universitas 21 organization: 2. Telemedicine and underserved populations. *J Telemed Telecare, 11*, 221–224.

KEY TERMS

Digital Divide: The explosive developments of Information and Communication Technology (ICT) in the last 10 years have led to a digital divide of the world. Those regions not having access to modern ICT will also lag behind in other fields.

E-Learning: The term e-learning is largely used for continuous medical education, both remote and local. Elementary basic information is usually given for free, whereas more structured and complete information is charged.

EMISPHER: Euro-Mediterranean Internet-Satellite Platform for Health, Medical Education and Research. The main objective of the EMISPHER project is to establish a telemedicine service platform integrating Internet-based and highly interactive applications via satellite to foster networking among all health care professionals in Euro-Mediterranean countries.

Evidence-Based Medicine: Selection of preventive, diagnostic, or therapeutic methods on the basis of scientifically based empirical evidence.

Medical Assistance: Collaborative access to medical images, electronic patient files, and so forth to assist treatment of patients who are abroad for holidays or work.

Quality of Service: The optimal design of networks for telemedicine should meet technical aspects such as transmission bandwidth, delay, jitter, data loss, and so forth (quality of service).

Real-Time Telemedicine: Telemedical applications with real-time transmission of live medical video allowing to work interactively.

Virtual Hospital: The objective needs of the heterogeneous partners in a telemedicine network can only be joined by a real integration of the various platforms and services. A virtual combination of applications serves as the basic idea for the Virtual Hospital.

Visual Methods for Analyzing Human Health Data

C. Tominski
Institute for Computer Science, University of Rostock, Germany

P. Schulze-Wollgast
Institute for Computer Science, University of Rostock, Germany

H. Schumann
Institute for Computer Science, University of Rostock, Germany

INTRODUCTION

Day by day, large volumes of health-related data are collected by physicians, health insurance companies, and public authorities. These data are potentially useful to understand the history, monitor the present, and predict the future of the health situation in order to ensure a high level of human health protection. To take advantage of this potential, it is necessary to analyze the data. However, this is a demanding task when facing constantly growing volumes of data.

One approach to tackle the analysis of human health data is the application of **visual methods**. In recent years, visualization of data has become a commonly accepted and widely used tool for the extraction of relevant information from arbitrary data. In many cases, a better insight into complex processes and phenomena can be gained by means of visual representation.

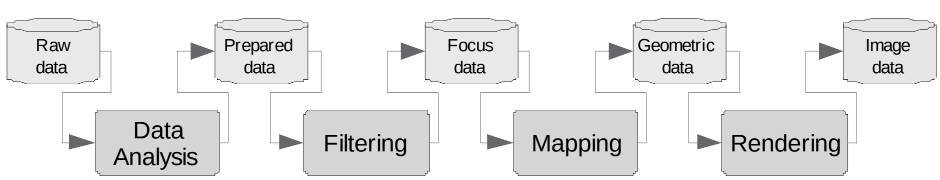
This chapter focuses on the **visual analysis** of human health data that describe the number of cases of different diagnoses in a spatial and temporal frame of reference. To build a common basis for the later description of different visualization methods, basic concepts of visualization as well as an abstract data model are illustrated in Background. In the main part of this chapter, we describe the visualization of human health data at various levels (see Visualizing Human Health Data). Whereas basic visual methods for representing human health data are presented only briefly, the visualization of data with respect to space and time is described in more detail. This chapter concludes with remarks on future work and trends in Future Trends and a brief summary of the key issues described in this article (see Conclusion).

BACKGROUND

Thanks to the capabilities of the human visual system, visualization is a promising tool to analyze larger volumes of data. If visualization is done properly, relevant information can be perceived intuitively, and the underlying data can be understood more easily. By proper visualization, it is meant that a visual representation has to be *expressive*, *effective*, and *appropriate*. Expressiveness relates to the requirement that all relevant information must be expressed in a visualization. Effectiveness depends on the degree to which a visualization supports easy and intuitive interpretation of the visualized facts. A visual representation is appropriate if gained benefit and required effort are balanced with respect to the task at hand.

Technically, the visualization process is implemented in four main steps: *data analysis*, *filtering*, *mapping*, and *rendering* (see Figure 1). They make up the **visualization pipeline** (dos Santos & Brodlie,

Figure 1. The visualization pipeline (dos Santos & Brodlie, 2004)

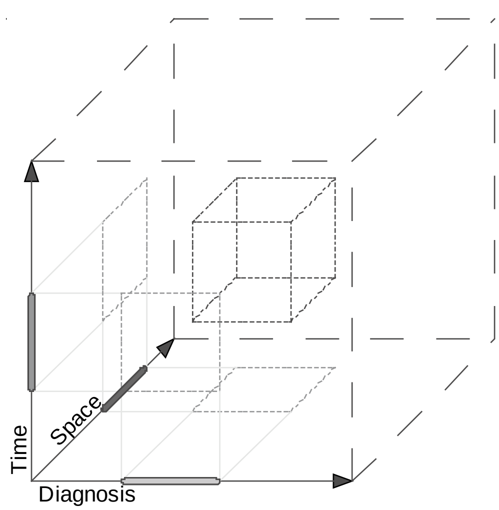


2004). To create a visual representation, a dataset is processed as follows. In the data analysis step, data are prepared for visualization (e.g., by applying a smoothing filter, interpolating missing values, or correcting erroneous measurements). The filtering step selects the data portions to be visualized (denoted as focus data). In the mapping step, focus data are mapped to geometric primitives (e.g., points, lines) and their attributes (e.g., color, position, size). The mapping step is the most critical one for achieving expressiveness and effectiveness, and hence, it is the most interesting one to visualization designers. Finally, geometric data are transformed to visual representations (e.g., images or animations).

Visualization aims at gaining insight by visually representing data. This implies that data characteristics are fundamental for any visualization. For this reason, research on visualization techniques, concepts, or methodologies must be grounded on a description of the addressed kind of data.

In this chapter, we address **human health data** that describe the number of cases of various diagnoses collected in a spatio-temporal frame of reference. Our data describe on a daily basis how many cases occur per diagnosis and per geographical region. From an abstract point of view, the data can be modeled as a data-cube (see Figure 2) that is constituted of three dimensions: time, space, and diagnosis. All these dimensions are of a hierarchical nature. Time uses days, months, quarters, and years; the spatial dimension comprises different administrative partitions of space (i.e., federal state, counties, and municipalities). The diagnoses are linked to the International Classification of Diseases (ICD10), which is hierarchical by definition. By relying on the data-cube model, it is relatively easy to reduce the volume of data to be visualized (e.g., by selecting only subranges of the dimensions [see Figure 2], or by using different levels of hierarchical abstraction). By this, only relevant data have to be extracted from the database. We will see in the next section that different visualization techniques vary in their usefulness for analyzing human health data depending on which dimensions of the data-cube are considered to what extent.

Figure 2. Different dimensions of health data modeled as a data-cube



VISUALIZING HUMAN HEALTH DATA

In the previous section, we indicated that human health data can be analyzed quite efficiently using visual methods. In this section, we suggest concrete **visualization techniques**. Depending on which dimensions of the described data-cube are addressed, different ways of representing human health data are possible:

- Multivariate representation of diagnoses
- Representation of diagnoses with respect to space
- Representation of diagnoses with respect to time
- Representation of diagnoses in time and space

Multivariate Representation of Diagnoses

In this first case, human health data are interpreted without spatio-temporal dependencies (i.e., only the diagnosis dimension of the data-cube is considered). In addition to diagnoses, it is also possible to enhance this abstract interpretation with derived statistics (e.g., maximum, minimum, average, and mean). The advantage of this interpretation is that classic visualization techniques can be applied to represent diagnoses and/or derived statistical characteristics visually. Simple diagram techniques (Harris, 1999) can be used to visualize frequencies of diagnoses (e.g., histograms for absolute frequency, pie charts for relative frequency). More sophisticated multivariate techniques such as Scatter Plot Matrices (Cleveland, 1993) or Parallel Coordinates (Inselberg, 1998) can be used to compare several diagnoses to reveal correlations among them (see Figure 3).

Although these first visual representations are quite abstract, they are nonetheless an initial step toward insight into the underlying data. The general number of cases as well as outliers can be recognized. The drawback is that the data's frame of reference is not considered (e.g., where is a peak in the number of influenza cases?).

Representing Diagnoses in Space

A representation of the spatial frame of reference is necessary to allow an easy interpretation of data features in their geographical context. Maps have been used for ages to illustrate information with respect to geographic realties. Maps have the advantage that they enable the representation of data with respect to natural or manmade spatial phenomena like continents or states, which in our case are given as irregular-shaped regions at different spatial granularities.

If only one diagnosis has to be represented on a map, Choropleth Maps are commonly applied. In Choropleth Maps, each geographical region is color-coded according to the number of cases that occurred for the respective region (see Figure 4a). If multiple diagnoses must be visualized, icon-based techniques are a good choice. Icons are small graphical primitives capable of encoding multiple attributes simultaneously. Furthermore, icons can be placed on maps to communicate spatial dependencies of diagnoses. Figure 4b shows icons representing three different diagnoses for user-selected regions. A challenge that arises when using icons on maps is to avoid occlusion of icons. Particularly for areas with many small regions, this is a demanding task (Fuchs & Schumann, 2004).

Representing Diagnoses in Time

To gain better understanding of the data, it makes sense to consider temporal dependencies. Even though visualization of time-related data is a challenge in its own right (Aigner, 2006), a compact overview of

Figure 3. Health data represented as (a) Scatter Plot Matrix and (b) Parrallel Coordinates

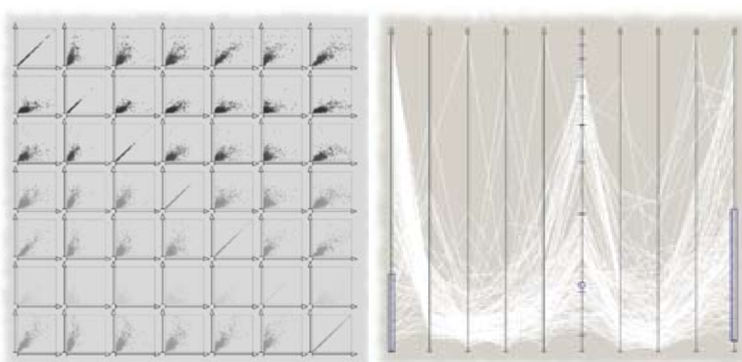
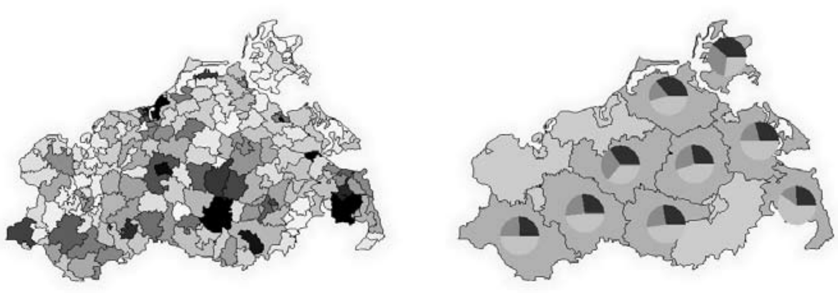


Figure 4. Health data represented via Chorpleth Map (a) and via icons positioned on a map (b)



(a) **(b)**

visualization techniques suitable for health data will be given.

A differentiation that must be made when visualizing temporal data regards the elements used for modeling the temporal dimension. Basically, a temporal domain is made up of either time points or time intervals[1]. Frank (1998) suggests a further categorization of time that regards the "shape" of the time axis. Frank distinguishes between linear, cyclic, or branching time. Whereas linear time considers a straight time axis from the past to the future, a cyclic time axis comprises recurring temporal elements (e.g., seasons of the year). Branching time axes are used to describe alternative scenarios of which only one will actually happen. These distinctions imply that, for instance, diagnoses that show cyclically recurring patterns (e.g., influenza) should be represented differently from illnesses that do not show such behavior (e.g., bone illnesses). Whereas techniques for visualizing health data with respect to time intervals or branching time are rare (e.g., (Plaisant, Mushlin, Snyder, Li, Heller & Shneiderman, 1998)), the visual representation with respect to time points given on linear or cyclic time axes is considered in various visualization approaches. From simple yet expressive diagram techniques (Harris, 1999) to more sophisticated techniques such as Spiral Displays (Weber, Alexa & Müller, 2001), TimeSearcher (Hochheiser, 2002) or parallel 3D bar charts (Chitarro, Combi & Trapasso, 2003), a broad range of tools for visually analyzing human health data is available.

A natural goal when representing human health data with respect to time is to reveal relations, and by this, to find correlations in the underlying dataset. How this goal can be accomplished is illustrated by the example of two visualization techniques: time diagrams and spiral displays. A time diagram is useful to compare different diagnoses with respect to a linear time axis (see Figure 5a). A simple line-plot is used to facilitate an intuitive interpretation of the evolution of diagnoses in time. However, for detecting periodic

Figure 5. A single diagnosis represented a time diagram (a) and as a spiral display (b)

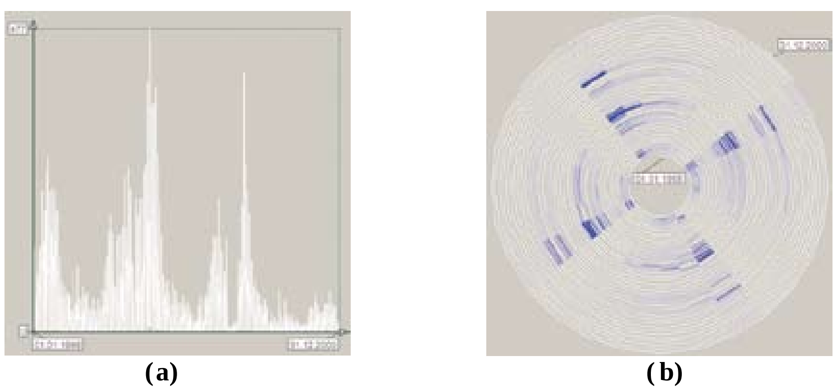
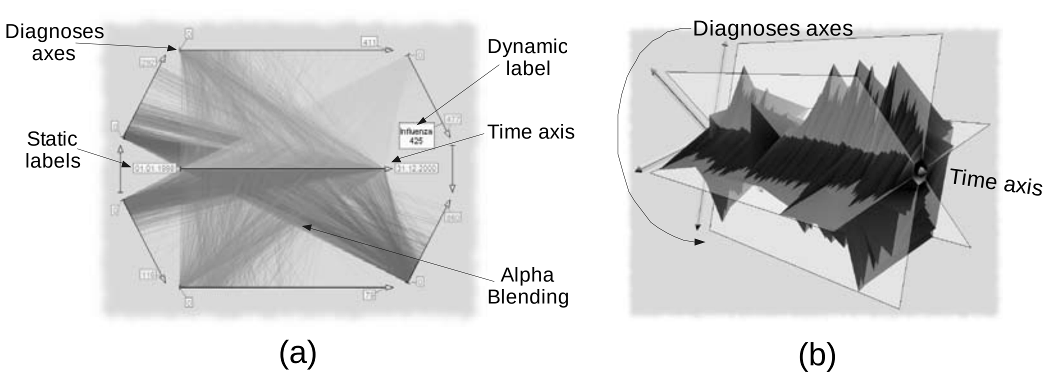


Figure 6. A TimeWheel (a) or a Kiviat Tube (b) can be used to visualize multiple diagnoses with respect to time



patterns, Spiral Displays are better suited. Spiral Displays represent the time domain as a spiral-shaped time axis (see Figure 5b). To enable the detection of periodic patterns of different length, Spiral Displays allow users to interactively adjust the number of time points encoded per spiral cycle.

Since the two illustrated visualizations and also most of the techniques known in the literature allow the visualization of only a few diagnoses, there is the need to restrict the visualization to important portions of the data. It makes sense to apply statistical analyses in the filtering step of the visualization pipeline to predetermine diagnoses that potentially correlate, and choose highly correlated diagnoses for visualization.

Techniques that are suited for representing multiple diagnoses with respect to a linear time axis are, for instance, the TimeWheel and the KiviatTube (Tominski et al., 2005). The TimeWheel comprises a central, horizontally aligned time axis around which several axes (representing diagnoses) are arranged (see Figure 6a). To represent data values in a TimeWheel, lines descend from the time axis to all other axes. Clearly speaking, for each time step, individual line segments establish direct visual connections between time and all corresponding time-dependent values. The TimeWheel enables users to easily discern the number of cases of the represented diagnoses with respect to a particular time step. The TimeWheel also allows determining where in time certain numbers were measured. Correlations among the diagnoses can be revealed using the KiviatTube. It is constituted by a central time axis and several axes (representing diagnoses) that emanate perpendicular to the time axis. This arrangement of axes makes it possible to render a dataset as a closed three-dimensional surface (see Figure 6b). The KiviatTube can be understood as a data representation that has been extruded along the time axis in 3D (similar to expanding an accordion). As such, the Kiviat Tube encodes not only the evolution of diagnoses over time, but also correlations among diagnoses.

Representing Diagnoses in Time and Space

Even though it is sufficient in certain analysis scenarios to consider time and space separately, it is more often the case that both spatial and temporal aspects have to be visualized. To consider all three dimensions of the data-cube (see Background), three basic tasks must be addressed:

1. Represent the spatial frame of reference as map
2. Create visual representations of the temporal aspects of the data
3. Embed representations of time (2) into the representations of space (1)

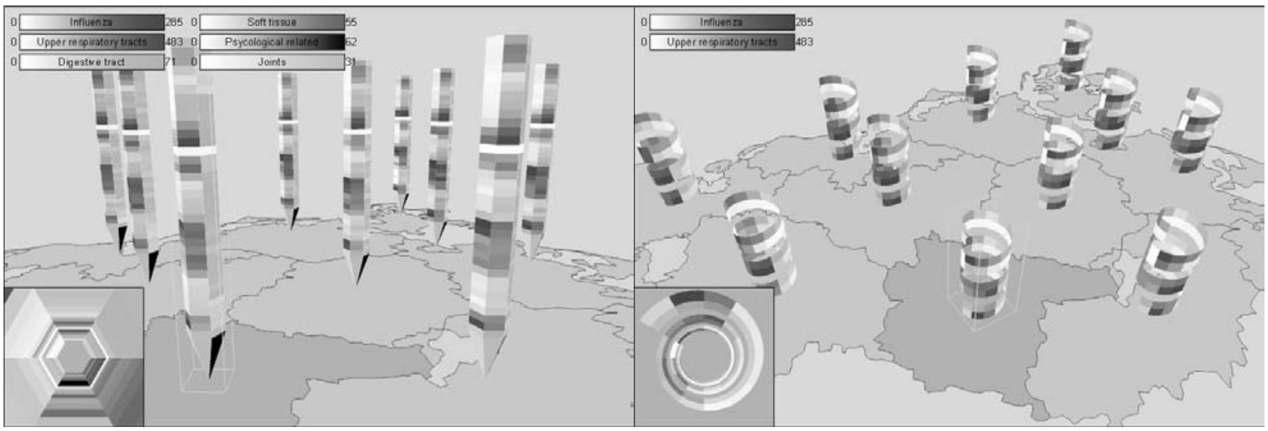
To represent the spatial frame of reference, the two-dimensional map (as introduced in Representing Diagnoses in Space) is represented in a 3D presentation space in which the map can be zoomed and rotated using common 3D interaction. Users that prefer a two-dimensional map representation can switch easily to a respective map projection.

To visualize temporal aspects, pencil icons are used for linear time axes, and helix icons are applied for cyclic time (Tominski et al., 2005):

- **Pencil icons:** Since everybody is familiar with the shape of a pencil, it is a very useful visual metaphor. The natural shape of a pencil provides multiple faces that evolve from a common tip. This shape excellently serves as a 3D icon for visualizing multiple diagnoses with dependencies on linear time. To map data onto a pencil icon, linear time is encoded along the pencil's faces starting from the common tip. Each face of a pencil is used to color-code one diagnosis (cp. Figure 7a). The number of faces of a pencil can be adjusted according to the number of diagnoses to be visualized.
- **Helix icons:** A spiral helix provides a geometric shape that allows an emphasis of the cyclic character of diagnoses. In order to construct a helix icon, a ribbon is created: For each time step the ribbon extends in angle and height depending on the number of temporal primitives per cycle and the number of cyclic passes. Again, color-coding is used to encode data values along the ribbon (cp. Figure 7b). In order to represent multiple diagnoses, the ribbon can be divided into narrower subribbons.

The described pencil and helix icons can be easily embedded into a 3D map display. This can be realized by positioning 3D icons at the centroids of geographical regions and aligning them with the z-axis of the 3D

Figure 7. Visualizing monthly health data by means of 3D icons on a map: (a) Pencil icons representing the number of cases of six diagnoses; some diagnoses show a certain pattern over time; (b) Helix icons clearly reveal the cyclic characteristic of two selected diagnoses. Additional "tunnel views" mitigate, for a selected icon, the problem of hidden information



(a) (b)

presentation space. By doing so, the representation of the temporal dependencies is shifted from the map to a dedicated dimension in the 3D presentation space. However, the embedding in 3D involves new problems compared to a 2D representation:

- Undesired changes of the icon view upon user interaction
- Loss of information due to icon occlusion and hidden surfaces

Applying all interactions allowed in the 3D presentation space (i.e., rotation, translation, scaling) to both the map and the icons is disadvantageous, since it implies visual inconsistencies during the visual analysis. If users rotate the map while analyzing diagnoses represented by a pencil icon, the faces of the pencil are rotated as well (i.e., the view of diagnoses changes). An alternative is to unlink map interactions and icon interactions on demand. This can be achieved by aligning the icons with respect to the user's current view into the 3D scene. This ensures that the icon view is preserved during map navigation. Users are allowed to change the icon view via separate icon rotation.

The second problem that must be addressed is loss of information due to occlusion of icons and on back faces of 3D icons. To alleviate this problem, the crude procedure of positioning icons at the areas' centroids must be improved. An iterative approach that calculates occlusion conflicts and alters the positions of icons locally can help to reduce occlusion. By applying an iterative algorithm that runs until no conflicts occur or until a maximum number of iterations has been exceeded, enhanced icon positions can be achieved, and interactive frame rates are ensured. To mitigate the problems imposed by information lost on back faces, it makes sense to provide the possibility to switch to special designated viewpoints. These viewpoints are located directly underneath each icon, and the view directions are aligned with the icons' axes. Switching to one of these viewpoints results in a "tunnel view" that reveals all data values represented by a selected 3D icon. Figure 7 shows how tunnel views support visual analysis in a 3D presentation space.

These enhanced viewing and interaction techniques in combination with common 3D interactions provide a rich basis for a visual exploration of multiple time and space dependent diagnoses. Furthermore, it is possible to emphasize linear or cyclic characteristics of data by using dedicated 3D pencil or helix icons, respectively.

FUTURE TRENDS

We presented a variety of visual methods for analyzing human health data. However, in light of large datasets, visualizing all data in a comprehensible manner without burying possibly important information becomes more and more challenging. This challenge can be dealt with by conducting additional data analysis steps. Statistical and *data mining* methods such as Principle Component Analysis or clustering are helpful analytical tools to support the identification of what is important in human health data. Although the analysis step is indispensable, it is yet mainly data-driven. In the future, the interests of users should also be considered. This will allow for providing automatically adjusted visualizations that emphasize relevant parts or hiding of nonrelevant data. A further particularly challenging problem is to find new ways to describe tasks of the visual exploration process and to automatically adapt the whole analysis procedure according to the users' tasks at hand. This also includes providing specific interaction functions for investigating health data.

CONCLUSION

The analysis of human health data can be supported by interactive visual interfaces and appropriate visualization techniques. To achieve the main goal of visualization—gaining deeper insights into larger datasets—it is crucial to consider the characteristics of the data and the needs of the analyst. In this chapter, we investigated the role of visualization in the context of human health data. We elaborated on choosing visualization techniques properly with respect to characteristics of the data. In particular, we described techniques for visually analyzing multivariate human health data in time and space. Although the explained techniques are promising tools to support better health care, more work (as indicated in Future Trends) has to be done to fully exploit the usefulness of visual analysis methods.

REFERENCES

Aigner, W. (2006). *Visualizing time and time-oriented information: Challenges and conceptual design* [doctoral thesis]. Vienna: Vienna University of Technology.

Chittaro, L., Combi, C., & Trapasso, G. (2003). Data mining on temporal data: A visual approach and its clinical application to hemodialysis. *Journal of Visual Languages and Computing, 14*(6).

Cleveland, W.S. (1993). *Visualizing data*. Summit, NJ: Hobart Press.

dos Santos, S., & Brodlie, K. (2004). Gaining understanding of multivariate and multidimensional data through visualization. *Computers & Graphics, 28*(3).

Frank, A.U. (1998). Different types of "times" in GIS. In M.J. Egenhofer, & R.G. Golledge (Eds), *Spatial and temporal reasoning in geographic information Systems*. New York: Oxford University Press.

Fuchs, G., & Schumann, H. (2004). *Intelligent icon positioning for interactive map-based visualization systems*. Proceedings of the 15th IRMA International Conference.

Hajnicz, E. (1996). *Time structures: Formal description and algorithmic representation*. Number 1047, Lecture Notes in Computer Science. Berlin: Springer-Verlag.

Harris, R.L. (1999). *Information graphics: A comprehensive illustrated reference*. Atlanta, GA: Management Graphics.

Hochheiser, H. (2003). *Interactive graphical querying of time series and linear sequence data sets* [doctoral thesis]. University of Maryland.

Inselberg, A. (1998). A survey of parallel coordinates. In H.-C. Hege, & K. Polthier (Eds.), *Mathematical visualization*. Berlin: Springer-Verlag.

Müller, W., Nocke, T., & Schumann, H. (2006). *Enhancing the visualization process with principal component analysis to support the exploration of trends*. Proceedings of the Asia Pacific Symposium on Information Visualization.

Plaisant, C., Mushlin, R., Snyder, A., Li, J., Heller, D., & Shneiderman, B. (1998). *LifeLines: Using visualization to enhance navigation and analysis of patient records*. Proceedings of the American Medical Informatic Association Annual Fall Symposium.

Schulze-Wolgast, P., Tominski, C., & Schumann, H. (2003). *Visual analysis of human health data*. Proceedings of the 14th IRMA International Conference.

Tominski, C., Abello, J., & Schumann, H. (2005). *Interactive poster: 3D axes-based visualizations for time series data*. Proceedings of the IEEE Symposium on Information Visualization.

Tominski, C., Schulze-Wollgast, P., & Schumann, H. (2005). *3D information visualization for time dependent data on maps*. Proceedings of the International Conference on Visualisation.

Weber, M., Alexa, M., & Müller, W. (2001). *Visualizing time-series on spirals*. Proceedings of the IEEE Symposium on Information Visualization.

KEY TERMS

Appropriateness: A visual representation is appropriate if its benefit and creation effort are balanced.

Data Cube: The dimensions of a dataset can be modeled as a hyper-cube, which allows an easy selection of relevant parts of the data.

Effectiveness: A visual representation is effective if it has been generated in accordance with the capabilities of the human visual system.

Expressiveness: A visual representation is expressive if it communicates all information that is relevant with respect to the current analysis task.

Interaction: Denotes the use of specific techniques to adjust visual representations according to the task at hand.

Multivariate Data: Data that contain multiple attributes measured in some observations space are denoted multivariate data.

Visualization: A visual approach to gaining insight into data.

Visualization Pipeline: Describes data analysis, filtering, mapping, and rendering as steps of visualization.

ENDNOTE

1 Since it can be assumed that the reader has a natural understanding of the terms “time point” and “time interval,” a formal description will not be presented here; interested readers are referred to Hajnicz (1996) or Aigner (2006).

In Vivo Near Infrared Techniques for Protein Drug Development

Yueqing Gu
China Pharmaceutical University, China

Zhiyu Qian
Nanjing University of Aeronautics and Astronautics, China

Huimin Qian
China Pharmaceutical University, China

Chunsheng Fang
China Pharmaceutical University, China

Yulin Song
Memorial Sloan-Kettering Cancer Center, USA

INTRODUCTION

Near infrared (NIR) light (700 ~ 900 nm) possesses the capability of penetrating living tissues several centimeters due to the low absorbance of tissue intrinsic chromophores such as oxy- and deoxy-hemoglobin (the main absorber of visible light), melanin, water, and lipid (the principal absorber of infrared light). Featured with the deeper tissue penetration as well as nonionizing and nonradioactive, NIR light attracts extensive attentions on the development of noninvasive techniques for *in vivo* real time monitoring/tracing of biological signals in living tissues. Hitherto, *NIR techniques* have permeated to almost all aspects of health care, such as diagnosing disease (Nahum, Skippen, Gagnon, Macnab, & Skarsgard, 2006), designing the targeted molecular or drug carrier (Hsu et al., 2006), monitoring the response to therapeutic treatment (Tachtsidis et al., 2007), evaluating the rehabilitation, and so on. With the rapid development of various NIR techniques and more cooperation with clinic studies, more potential applications in health care will be exploited in the near future.

Characterized with the virtue of high physiological activity, specificity, and low toxicity, *protein drugs* are becoming mainstream therapeutic agents and constitute a substantial portion of the compounds under preclinical and clinical development in the global pharmaceutical industry (Baumann, 2006). And the *in vivo* evaluation of drug properties, from animal models to human subjects, becomes a crucial in the *protein drug* development. It involves drug screening, drug delivery, drug biodistribution, pharmacokinetics, pharmacodynamics, and so on. Each of these processes requires *quantitative* monitoring of *drug concentrations* in the specific organ/tissue or *qualitative imaging* of *drug distribution* in whole subjects. Thus, a number of analytical techniques were exploited to analyze the *protein drug* process within the experimental subjects. Immunoassay, bioassay, and isotope labeling are the most commonly used methods for quantification of macromolecule in biological fluid (Marshall, Macintyre, James, Krams, & Jonsson, 2006). Immunoassay is based on the specific antibody–antigen reaction. The response signal is generated from a label (e.g., enzymatic, fluorescent, and radio isotopic) attached to either antigen or antibody. This method, including radioimmunoassay (RIA), immunoradiometric assays (IRMA), and enzyme-linked immunosorbent assays (ELISA), is rapid and sensitive as well as economical, but it is heavily affected by a variety of endogenous and exogenous substances. Bioassay provides an indirect method by using living system to measure the biological activity of a drug. This method is expensive, time consuming, and lacks specificity and sensitivity. Isotope labeling is an alternative approach to *in vivo* quantification of protein drug (Liu, Dreher, Chow, Zalutsky, & Chilkoti, 2006). Although it is sensitive, some disadvantages regarding the complex preparation and radioactivity limited its broad application. Beside the aforementioned analytical methods, techniques such as chromatography and electrophoresis are usually used for *protein drug measurements*. However, all the listed techniques, except isotope labeling, are unable to perform the *in vivo* measurement of *protein drug* in biological matrices. They require the tedious and intricate process of sample preparation such as sample collection and purification. Furthermore, the clearance of protein drugs from the specific tissue/organ is quick and the above analytical techniques cannot catch up with its dynamic changes.

Therefore, a noninvasive *in vivo* real time monitoring modality for protein drugs in biological matrices will timely meet the needs and greatly expedite the step of *protein drug* development.

Near infrared techniques contributed considerably to the development of protein drugs by performing noninvasive *in vivo* real time *measurements* on experimental subjects. In this chapter, we will first present the fundamental theory of NIR techniques, followed by the summary of NIR *fluorescence probes*. Emphasis will be placed on the practical applications of NIR techniques on drug development. Future trends and conclusions will be given at the end.

BACKGROUND

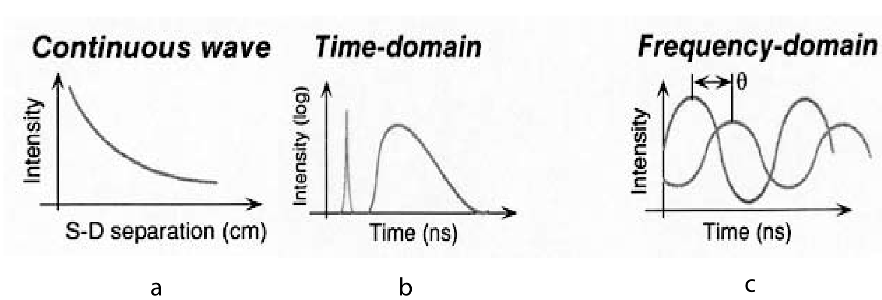
Near Infrared Techniques: Spectroscopy and Imaging

Near infrared light (700 to 900 nm), where photon transport in tissue is dominated by scattering rather than absorption, has the maximal penetration depth of several centimeters, therefore enabling *in vivo* sampling of large tissue volumes, for example, breast, brain, skeletal muscles, or tumors. The absorptions of NIR light by endogenous chromospheres, that is, oxygenated and deoxygenated hemoglobins, and exogenous contrast agents have been used to determine important physiological characteristics of tissues.

In the last decade, extensive investigations in noninvasive NIR techniques have been conducted for both *quantitative* and *qualitative* determination of biological signal in tissues. Based on the profile of the incident light, three main near infrared systems, that is, time-domain (TD), frequency-domain (FD), and continuous wave (CW) photon migration system, have been developed for various *measurements*. Figure 1 illustrates the profiles of the incident light (red line) and tissue diffused light (blue line) for the three systems.

Usually, CW techniques employ illumination with a constant intensity. The diffused light attenuates exponentially with the source-detector (S-D) separation, section (a) of Figure 1, and with the depth of the tissue according to the tissue optical properties. Although having limited penetration depth and only intensity retrieved, CW system is the most easily implemented system and the most widely used in various *measurements* on small *animal models*. In contrast to CW techniques, TD approaches employ a source with a femto-picosecond pulse of light, which is broadened and attenuated as it propagates through the tissue, see section (b) of Figure 1. The diffused light intensity, transient time, and pulse shape are recorded to determine the tissue optical properties. The measurement from a TD system is more accurate and usually used as calibration for other newly developed systems. However, its high cost has limited its wide application. FD techniques offer a compromised approach to the determination of tissue optical properties. The incident light on an FD system is modulated at radiofrequency range (namely, 70 MHz—several hundred MHz). The diffused photon wave propagates through tissue and becomes amplitude-attenuated and phase-shifted relative to the incident light. After demodulation, the changes in amplitude and phase are retrieved for the determination of tissue optical properties. Various forms of heterodyne or homodyne demodulation systems have been utilized to implement FD system. Because the detected amplitude and phase are insensitive to the ambient light, FD system enables greater sensitivity and

Figure 1. Profiles of the incident light (red line) and the diffused light (blue line) for the continuous wave system, the time-domain system, and the frequency-domain system



penetration depth than CW measurements. For all the above systems, multiple NIR wavelengths were usually applied to distinguish the different chromophores in tissues based on the modified Beer-Lambert's Law. The propagation of NIR fluorescence follows the same rule as the excitation light as long as the lifetime of the dye is less than or comparable to the timescale of excitation light propagation.

No matter which incident light applied, the NIR system can be implemented as *spectroscopy* and *imaging*. Figure 2 illustrates the schematic diagram of a near infrared spectroscopy (NIRS) system. The processing circuit is matched to the light sources via a coupling mechanism depending on the type of modulation used. The geometry of delivering and collecting fiber bundles is flexible and can be adjusted as reflectance, transmittance or fluorescence modes according to practical applications.

NIR spectroscopy is usually used for quantitative determination of chromophores concentration in specific region of the subjects. Oxygenated and deoxygenated hemoglobins are the most common physiological parameters to be measured by NIRS for evaluation of vascular hemodynamic properties (Gu, Bourke, Kim, Constantinescu, Mason, & Liu, 2003; Liu, Song, Worden, Jiang, Constantinescu, & Mason, 2000) and disease diagnosis. However, efforts in applying NIRS to *in vivo* monitoring of *drug concentration* are very limited and preliminary, which holds great promising in drug development. While intrinsic chromophores absorption yields significant information on tissue characteristics, extrinsic fluorescence markers attract tremendous attention for *in vivo* monitoring of biological signal in tissue, which is particularly useful in drug development.

The need of tissue observation facilitates the rapid development of near infrared *imaging*. And the extrinsic fluorescence markers bring NIR imaging a prodigious achievement in all the aspects of health care. Currently, reflectance mode and tomography are the two most common NIR imaging approaches used for observing tissue properties. The schematic diagrams are shown in Figure 3. NIR reflectance *imaging* is usually equipped with a continual wave light source (either a laser at an appropriate wavelength or a white light furnished with a bandpass filter) and a high sensitivity CCD camera with a filter for capturing the emitted fluorescence from the tissue. It is currently the typical method for accessing the distribution of fluorescent marker in biological matrices, including in drug development. Together with biological marker, NIR reflectance *imaging* offers additional insights into the determinants of disease progression and therapeutic response. Although NIR reflectance imaging is an ideal tool for imaging and screening of molecular events close to the surface, it has fundamental limitations on the penetration depth and quantitative measurement. To resolve and quantify fluorochromes deep inside the tissue, NIR tomographic methods, section (b) of Figure 3, are developed to fill the gap. Except for the hardware furnished with the multiple source-detector pairs, mathematical algorithms based on photon migration are crucial for reconstructing fluorochromes *imaging* inside deep tissue with better spatial resolution. Up until now, various tomographic *imaging* systems, that is, frequency-domain, time-domain, and continuous-wave imaging, have been implemented and applied mostly in brain function and breast cancer study. There is little report on quantification of drug concentrations from NIR imaging approaches.

Figure 2. Schematic diagram of a near infrared spectroscopy system. The dash lines represent the components in time-dependent system.

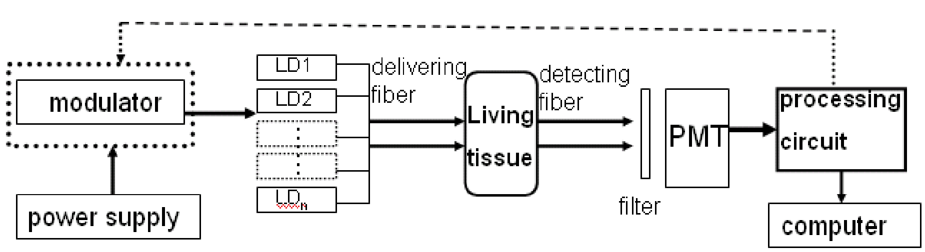
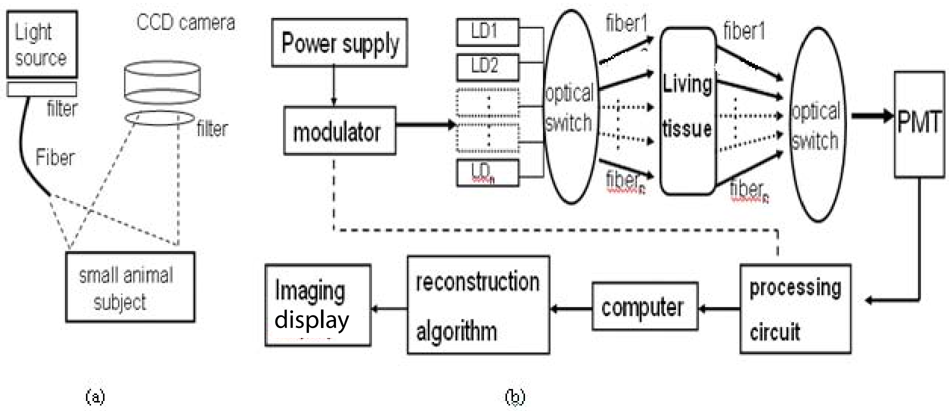


Figure 3. Schematic diagram of NIR reflectance imaging (a) and tomography (b)



Extrinsic NIR Fluorescent Markers: Organic and Inorganic Dyes

Due to the fact that most of the protein drugs lack unique optical properties within the near infrared range, the extrinsic NIR fluorescent labeling makes the *in vivo* study of *protein drug* possible. For a long time, extrinsic fluorescent agents were limited to organic fluorophores, most of which are pentamethine and hepamethines cyanines comprising bezoxazole, benzothaizole, indolyl, 2-quinoline or 4-quinoline. Recently, several new NIR dyes with functional groups, for instance, active ester, carboxyl group, sulfochloride, and so on, were synthesized for covalent conjugation to biological molecules. A part of them is listed in Table 1, including structures and fluorescent spectral properties. Their biodistributions, pharmacokinetics, and applications as extrinsic contrast agent have been summarized by Ballou, Ernst, and Waggoner (2005) and Licha and Olbrich (2005). These agents are always conjugated to biomolecules, including antibodies, peptides, saccharides, drug carriers, and so on for drug development.

Although conventional organic fluorophores have achieved great progress in the development of clinic medicine, their unfavorable characteristics, such as low quantum yield, difficult to manipulate the fluorescent wavelength, and susceptible to photobleaching, limited their applications in some cases. Inorganic fluorescent semiconductor nanocrystals, namely quantum dots (QDs), have the potential to overcome most of the drawbacks of organic fluorophores. Typically, QDs have a core/shell structure of 3-20 nm in diameter and may be multifunctionalized by many ligands and probes. With favorable properties, such as high absorbency, high quantum yield, narrow emission bands, large Stokes shifts, and high resistance to photobleaching, fluorescence quantum dots have been successfully applied to the study of cell and animal biology, immunofluorescence assays, DNA array technology, and so on. The red shift of fluorescence emission to NIR range is believed to have great promising for the *in vivo* study of drug development.

METHODOLOGY

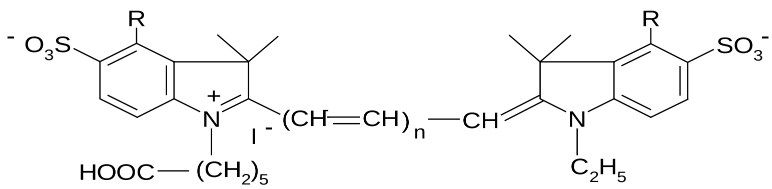
NIR Techniques Applied in Drug Development

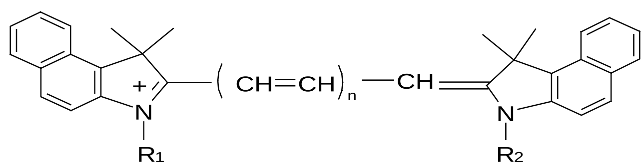
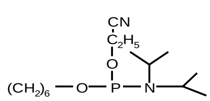
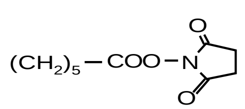
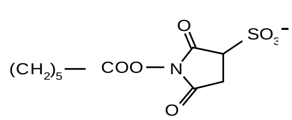
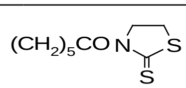
With the help of NIR fluorescence agents, *near infrared techniques* provide a valuable noninvasive tool to *in vivo* real time monitoring of the route/distribution or the efficacy of drugs on small *animal models*. In general, NIR *spectroscopy* offers a *quantitative* way to measure the *drug concentrations* in specific tissue/organs, while NIR imaging can image the distribution of drug/carrier in *animal models*. Although the drug quantification by NIR techniques is limited, the results from our group give us more confidence on the potential perspective of NIR techniques in drug developments.

Quantitative Measurements of Protein Drug Concentration for Pharmacokinetics

Dynamic *drug concentration* in *animal models* is the most crucial parameter in drug development, especially

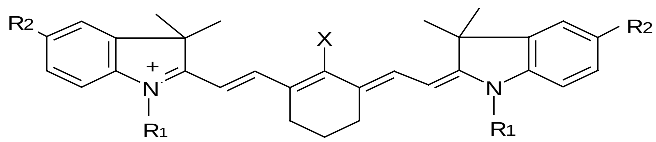
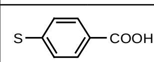
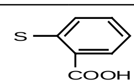
Table 1. Structures and optical properties of organic NIR fluorophores

						
NIR dye	R	n	$\lambda_{max}(nm)^{ab}$	$\lambda_{max}(nm)^{em}$	$\varepsilon \times 10^3$ $(cm^{-1}\ M^{-1})$	
Cy5	——	2	647	664	250	
Cy5.5	SO_3^-	2	674	689	250	
Cy7	——	3	747	776	250	

						
NIR dyes	R_1	R_2	n	λ_{max}^{ab} (nm)	λ_{max}^{em} (nm)	$\varepsilon \times 10^3$ $(cm^{-1}M^{-1})$
IRD 700		$(CH_2)_4$-SO_3^-	2	685	712	300
ICG	$(CH_2)_4$-SO_3^-	$(CH_2)_5$-SO_3^-	3	780	820	180
ICG-OSu	C_2H_5		3	——	——	——
ICG-sulfo-OSu	$(CH_2)_4$-SO_3^-		3	768	807	128
ICG-ATT		C_2H_5	3	765	830	——

continued on following page

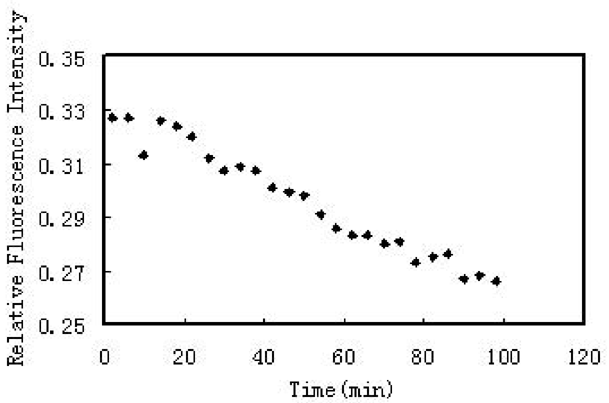
Table 1. continued

Cypate	C_2H_5COOH	C_2H_5COOH	3	787	816	224
						
NIR dyes	R_1	R_2	X	λ_{max}^{ab} (nm)	λ_{max}^{em} (nm)	$\varepsilon \times 10^3$ ($cm^{-1}M^{-1}$)
——	C_2H_5	——	Cl	771	——	——
——	C_2H_5	SO_3^-Na	Cl	776	——	130
NIR820	$(CH_2)_4$-SO_3^-	COOH	Cl	790	820	——
——	$(CH_2)_4$-SO_3^-	——	Cl	——	——	——
IRD40	$(CH_2)_4$-SO_3^-	——	O-C_6H_4-NCS	768	788	——
NN382	$(CH_2)_4$-SO_3^-	SO_3^-Na	O-C_6H_4-NCS	778	806	——
——	C_2H_5	——	S-CH_2COOH	792	823	289
——	C_2H_5	——	S-C_2H_5COOH	796	828	276
——	C_2H_5	——		806	841	275
——	C_2H_5	——		804	842	246

for the study of pharmacokinetics. As mentioned above, it's difficult to perform *in vivo* measurements for *protein drug* without extrinsic labeling. NIR dyes usually have a molecular weight of about 600 and are conjugated to protein drugs that have a larger molecular weight than those of dyes. The conjugation will not change the delivery properties of protein drugs. Therefore, the dye labeling enables the *in vivo* noninvasive measurements of *protein drug* dynamics. Furthermore, the dye labeling may enhance the drug to background ratio and, thus, improve the sensitivity and specificity of the measurements.

In our study, cypate, one of the NIR dyes provided by Dr. Achilefu (Ye, Bloch, Kao, & Achilefu, 2005), was covalently conjugated to a protein model drug L-asparaginase for *in vivo* concentration quantification. L-asparaginase is a remarkably effective therapy *protein drug* for those specific cases where blood cells become cancerous, such as in acute lymphoblastic leukemia. In the process of cypate conjugation to L-asparaginase, many factors including catalyzing system, *p*H condition, temperature, reacting duration, and feed ratio of different components were intensively investigated to achieve an optimal labeling result (Qian & Gu, 2006). In particular, the labeling efficiency, characterized by the dye-to-protein molecular ratio, and the enzyme activity of L-Asparaginase, evaluated by the Nessler's reagent photometric method, were carefully considered

Figure 4. Time profile of relative fluorescence intensity emitted from cypate-labeled L-Asparaginase in blood stream



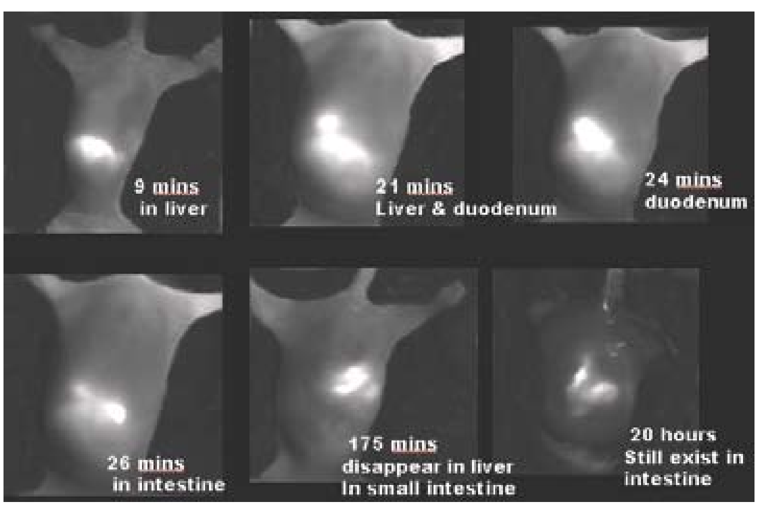
for higher monitoring sensitivity and better therapeutic efficacy.

The optimal conjugation of cypate-labeled L-Asparaginase, with higher labeling efficiency and enzyme activity, was administrated through a tail vein into the blood stream of mouse. The blood dynamic concentrations of L-Asparaginases were *in vivo* monitored by a NIR CW *spectroscopy* whose experimental setup is similar to that in Figure 2. Here, only fluorescence was directly recorded to quantify the cypate in blood. And the concentrations of L-Asparaginase were obtained by applying a precalibrated curve. A laser diode with a wavelength of 760 nm was used to excite the cypate to emit fluorescence at around 820 nm. A high pass filter (> 800 nm) was placed in front of PMT to block the incident light. The delivering and collecting fiber bundles were positioned in a perpendicular geometry in the leg area that contains abundant vasculature. Figure 4 shows the time profile of relative fluorescence intensity emitted from cypate-labeled L-Asparaginase in blood stream. The concentration of L-Asparaginase can be accurately quantified by the calibration. The experimental result from NIRS is consistent with the parallel measurement from radioisotopic labeling, which demonstrated the practicability and reliability of NIR *spectroscopy* as a quantitative monitoring tool. The detecting sensitivity on the dye-labeled drug is much higher than the measurement from the drug itself. The quantitative information of drug dynamics on small *animal models* speeds up greatly the pharmacokinetic study.

Qualitative Imaging of Protein Drug in Small Animal for Drug Development

NIR reflectance imaging is the most prevalent method applied to follow/trace the distribution of dye-labeled macromolecular compound in *animal models* for drug development. Similarly, we also utilized the NIR reflectance imaging system, section (b) of Figure 3, to successfully trace the route of cypate-labeled lysozyme in mouse, as shown in Figure 5. Cypate was covalently bound to lysozyme via two carboxyel groups with higher labeling efficiency and enzyme activity. The optimal conjugation of cypate-labeled lysozyme was

Figure 5. Dynamic route of cypate-labeled lysozyme in mouse. Model monitored by reflectance NIR imaging.



administrated through a tail vein into the blood stream of mouse. The distributions of lysozyme were qualitatively imaged at certain time intervals by reflectance NIR *imaging* furnished with a high pass filter before CCD camera. As demonstrated in Figure 5, cypate-labeled lysozyme was first accumulated in liver (9 minutes after administration) and then sequestered in duodenum and intestines. The back and forth moving of cypate-labeled lysozyme in duodenum and intestines was observed during the experiment. Most of the dye-labeled drug was cleared from liver at about 175 minutes, but still observed in intestines 20 hours after administration. These images revealed the clearance route of cypate-labeled lysozyme, providing important information for *protein drug* development and drug discovery.

FUTURE TRENDS

Despite advances in applying *in vivo* NIR techniques to facilitate the process of drug development, we are currently still limited in many aspects. First, most of the currently available NIR probes/contrast agents have a molecular weight of around 600. If the NIR agent is attached to a selected drug, it may change the delivery properties of the drug with a small molecular weight. This may be the reason why most of the studies on *in vivo* NIR techniques just involved in macromolecular compounds. Second, the quantitative measurement of *drug concentration* by NIR techniques is very limited and preliminary, which may be one of the major obstacles for the wider application of NIR techniques in drug development. Third, although the qualitative imaging of *drug distribution* by NIR imager in small *animal models* is very popular, there are few reports on the precise localization of drugs in a specific tissue. To accelerate the widespread utilization of NIR techniques in drug development, we need to overcome these limitations and explore more potential applications. Therefore, many investigators focus their research on the following directions:

1. Development of more and better NIR agents, with higher fluorescence quantum yield, more stable and less photobleaching, attachable not only to macromolecular drugs, but also to small molecular drugs; Synthesis of smart NIR agents with functional groups acting as targeted drug carrier (Ye, Bloch, Xu, & Achilefu, 2006); Conjugation of current NIR dyes to the existing probes such as radiotracers, MR markers to form multifunctional labeling agents for parallel measurements by using different modalities.
2. Improvement of high sensitive NIR systems. In particular, enabling NIR *spectroscopy* to be a commonly used quantitative monitoring tool for *drug concentration* determination in specific tissues at pico- to femto-mole accumulations, thus providing quantitative information for drug development.
3. Development of better algorithms for multiple source-detectors tomographic *imaging* system to precisely localize drug contents in deeper organ/tissue/cells, thus leading to quick drug screening and drug discovery.
4. Integration of NIR techniques with other mature modalities to perform parallel measurement for better understanding of drug-tissue interactions.

CONCLUSION

The emergence of NIR techniques offers great potential to expedite drug discovery and development. Although there are three typical NIR techniques based on the intensity profile of the incident light, spectroscopy and imaging are the most widely implemented systems for noninvasive *in vivo* real time monitoring of optical signals in biological matrices. Usually, NIR *spectroscopy* offers a quantitative way to determine the endogenous and exogenous chromophores in living body, while the NIR *imaging* provides a qualitative method to image the distributions of biomarkers in whole animal body. With the increasing availability of NIR fluorescent agents, NIR techniques have achieved exciting progress in many aspects of health care, particularly in drug development. Although having some drawbacks such as low quantum yield, inconvenience to manipulate the wavelength, susceptible to photobleaching, and organic NIR fluorescent dyes are still the most commonly used fluorescent agents for molecular imaging. Inorganic fluorescence agent, namely quantum dots, possesses some important advantages over organic agents and brings bright perspective in biological detection.

The results obtained from NIR *spectroscopy* and *imaging* in our experiments and from literatures demonstrated that NIR technique is a practical and reliable monitoring modality for *in vivo* tracing of *drug* *concentration* and distribution in small *animal models*. The widespread application of NIR technique will significantly expedite the step of drug development.

REFERENCES

Ballou, B., Ernst, L.A., & Waggoner, A.S. (2005). Fluorescence imaging of tumors in vivo. *Current Medicinal Chemistry, 12,* 795-805.

Baumann, A. (2006). Early development of therapeutic biologics-pharmacokinetics. *Current Drug Metabolism, 7,* 15-21.

Gu, Y.Q., Bourke, V., Kim, J.G., Constantinescu, A., Mason, R.P., & Liu H.L. (2003). Dynamic response of breast tumor oxygenation to hyperoxic respiratory challenge monitored with three oxygen-sensitive parameters. *Applied Optics, 42*(16), 2960-2967.

Hsu, A.R., Hou, L.C., Veeravagu, A., Greve, J.M., Vogel, H., Tse, V., et al. (2006). In vivo near-infrared fluorescence imaging of integrin alphavbeta3 in an orthotopic glioblastoma model. *Mol. Imaging Biol., 8*(6), 315-323.

Licha, K., & Olbrich, C. (2005). Optical imaging in drug discovery and diagnostic applications. *Advanced Drug Delivery Reviews, 57,* 1087-1108.

Liu, W., Dreher, M.R., Chow, D.C., Zalutsky, M.R., & Chilkoti, A. (2006). Tracking the in vivo fate of recombinant polypeptides by isotopic labeling. *Journal of Controlled Release, 114,* 184-192.

Liu, H., Song, Y., Worden, K.L., Jiang, X., Constantinescu, A., & Mason, R.P. (2000). Noninvasive investigation of blood oxygenation dynamics of tumors by near infrared spectroscopy. *Applied Optics, 39*(28), 5231-5243.

Marshall, S., Macintyre, F., James, I., Krams, M., & Jonsson, N. (2006). Role of mechanistically-based pharmacokinetic/pharmacodynamic models in drug development. *Clinic Pharmacokinetic, 45*(2), 177-197.

Nahum, E., Skippen, P.W., Gagnon, R.E., Macnab, A.J., & Skarsgard, E.D. (2006). Correlation of transcutaneous hepatic near-infrared spectroscopy readings with liver surface readings and perfusion parameters in a piglet endotoxemic shock model. *Liver Int., 26*(10), 1277-1282.

Qian, H.M., & Gu, Y.Q. (2006, October 16-18). Near-infrared fluorescent labeling of l-asparaginase. In *International Symposium on Biophotonics. Nanophotonics, and Metamaterials*, (pp. 124-128). Zhejiang, China.

Tachtsidis, I., Tisdall, M., Leung, T.S., Cooper, C.E., Delpy, D.T., Smith M.. et al. (2007). Investigation of in vivo measurement of cerebral cytochrome-c-oxidase redox changes using near-infrared spectroscopy in patients with orthostatic hypotension. *Physiol. Meas., 28*(2), 199-211.

Ye, Y., Bloch, S., Kao, J., & Achilefu, S. (2005). Multivalent carbocyanine molecular probes: Synthesis and applications. *Bioconjug. Chem., 16*(1), 51-61.

Ye, Y., Bloch, S., Xu, B., & Achilefu, S. (2006). Design, synthesis, and evaluation of near infrared fluorescent multimeric RGD peptides for targeting tumors. *J. Med. Chem., 49*(7), 2268-2275.

KEY TERMS

Drug Distribution: A different amount of drugs will be accumulated in different organs or tissues when administrated in the living subjects.

Fluorescence Probe: A specific agent that emits fluorescence with longer wavelength when excited by the light within its absorption range.

***In Vivo* Measurement:** Measurements are performed on the body of living subject (human or animal) without taking the sample out of the living subject.

Near Infrared: Electromagnetic wave whose wavelength ranges from 700 nm to 900 nm. These lights have higher penetration depth in living tissue due to low absorption coefficient to hemoglobin in blood.

Pharmacokinetics: A study for the dynamic change of drug concentration in blood. This parameter can help researchers and clinicians to understand the properties of the drug.

Protein Drug: A type of drugs made of protein. These drugs usually have large molecular weight with protein characteristics.

Volumetric Analysis and Modeling of the Heart Using Active Appearance Model

Sándor Miklós Szilágyi
Hungarian Science University of Transylvania, Romania

László Szilágyi
Hungarian Science University of Transylvania, Romania

Zoltán Benyó
Budapest University of Technology and Economics, Hungary

INTRODUCTION

Echocardiography is the fastest, least expensive, and least invasive heart imaging method. Accordingly, it is one of the most commonly used techniques to quantify the ventricular systolic function in patients. The examination is based on visual analysis of myocardial wall motion and deformation by an experienced and trained physiologist. This investigation is subjective, experience dependent, and the obtained results are only partially quantitative. The segmentation of the measured image sequences focuses on finding the exact boundaries of particular objects of interest, but it usually requires manual assistance.

Besides their several advantages, ultrasound images have the following drawbacks:

- They include not only the reflections from tissue transitions, but also several interference patterns (speckle noise). Consequently, tissues can hardly be distinguished by the intensity of their representing pixels.
- Image data highly depend on the position and angle of incidence of the ultrasound beam.
- A wide scale of imaging artifacts are frequently present, so still-frame images might contain only partial information.

In order to deal with these kinds of deficiencies, several automated segmentation techniques have been developed and reported. Geiser, Wilson, Wang, Conetta, Murphy, and Hutson (1998) proposed arc filtering for boundary detection, while Brotherton, Pollard, Simpson, and DeMaria (1994) gave a hierarchical fuzzy neural network solution. Dias and Leitão (1996) introduced an iterative multigrid dynamic programming technique based on Rayleigh distributed random variables and a probabilistic model formulated within Bayesian framework. Belohlavek, Manduca, Behrenbeck, Seward, and Greenleaf (1996) proposed the automated segmentation using a modified self-organizing map. Chalana, Linker, Haynor, and Kim (1996) traced the epi- and endocardial border using active contour models. In spite of their significant merits, these methods still neglect the following aspects:

- Sought boundaries are not always represented by the strongest edges.
- They use no a priori information concerning the allowable shapes and ranges of the segmented object.
- Segmented boundaries should be consistent with the cardiac cycle.

In the last decade, advances have been made in the content-based retrieval of medical images, such as extraction of boundaries of cardiac objects from echocardiography image sequences (Duncan & Ayache, 2000). Montagnat, Delingette, and Malandain (1999) used a two simplex mesh-based cylindrical deformable surface to produce time-continuous segmentation of 3-D sequences. Angelini, Laine, Takuma, Holmes, and Homma (2001) proposed a feature enhancement and noise suppression using a wavelet-like decomposition of the spatial frequency domain. A snake-based segmentation is carried out later on the denoised data.

Active appearance models (AAM), introduced by Cootes, Edwards, and Taylor (2001), are promising image segmentation tools that may provide solutions to most pending problems of echocardiography, as they rely on both shape and appearance (intensity and/or texture) information. Bosch et al. (2002) proposed a robust and time-continuous delineation of 2-D endocardial contours along a full cardiac cycle, using an extended AAM, trained on phase-normalized four-chamber sequences.

To understand the physiology and patho-physiology of the heart, not only the electrical activity and spatial distribution of its structures are important, but also their movement during normal and abnormal cardiac cycles. The ECG signal is measured simultaneously with echocardiography sequence recording in order to localize the investigated events. We developed an algorithm that reconstructs the heart wall boundaries and motion in order to determine the spatial and temporal cardiac activity.

In this article we present our algorithm that reconstructs the heart wall boundaries and motion in order to determine the spatial and temporal cardiac activity.

Several papers in the literature have already reported the usage of spatial AAM (Mitchell, Lelieveldt, Bosch, van der Geest, Reiber, & Sonka, 2002; Stegmann & Pedersen, 2005). The present work has the following contributions:

- Reported techniques classify ultrasound images only as belonging to systolic or diastolic interval. Our approach distinguishes normal and extra beats, and processes the corresponding images accordingly.
- ECG event classification makes possible the investigation of several pathological cases (e.g., volumetric effect of a given extra beat). Comparisons were made between normal and pathological cardiac cycles of the same patient.

MATERIALS AND METHODS

The echocardiography measurements were carried out as presented in the article with title "Echocardiographic Image Sequence Compression Based On Spatial Active Appearance Model" (see Figure 1).

The necessary minimal duration of the recorded image sequences was restricted by the semi-periodic behavior of the ECG signal. The spatial movement of the heart is constrained by the course of the depolarization repolarization cycle (Szilágyi, Benyó, & Szilágyi, 2005). For example, normal and ectopic beats imply different spatial heart movements. The studied ECG parameters, as presented in Figure 2, were: shape of QRS beat, QT, and RR distances. These parameters characterize the nature of a QRS complex, and were determined as presented in (Szilágyi, Benyó, & Szilágyi, 2003).

As the measurements were made with known spatial coordinates *x*, *y*, *z*, time *t*, and direction represented by angle α, the selected image sequences could be concomitantly processed with the measured ECG signal. All selected images that correspond to the same depolarization-repolarization phase and had almost the same ECG pattern (same QRS class, low QT, and RR difference) were included into calculations. Let $c(cl,t)$ be the clustering function that determines the appertainity coefficient to a cluster *cl* of a given ECG period measured at moment *t*. These clusters were determined with hermitage functions and self-organizing maps as presented in Lagerholm, Peterson, Braccini, Edenbrandt, and Sörnmo (2000).

The strength of the relation between a given wall segment and cluster *cl* is determined by function *f* with form:

$$f\left(x,y,z,cl,t\right)=\frac{c\left(cl,t\right)}{dist\left(x,y,z,t\right)}$$

where Box 1 represents the spatial distance at moment *t* of the sensor *s* from the studied wall piece *w*. The angle α of the sensor and the wall place at the image measured at moment *t* determinates the spatial coordinates of the muscle segment. In this way, each wall

Box 1.

$$dist(x,y,z,t)=\sqrt{\left(x_s\left(t\right)-x_w\left(t\right)\right)^2+\left(y_s\left(t\right)-y_w\left(t\right)\right)^2+\left(z_s\left(t\right)-z_w\left(t\right)\right)^2}$$

Figure 1. Various positions of the parallel cross sections and the orientation of the transducer: (a) Horizontal plane(s), (b) left rotated plane(s), (c) right rotated plane(s). Various positions of planes with common axis: (d) front view, (e) back view

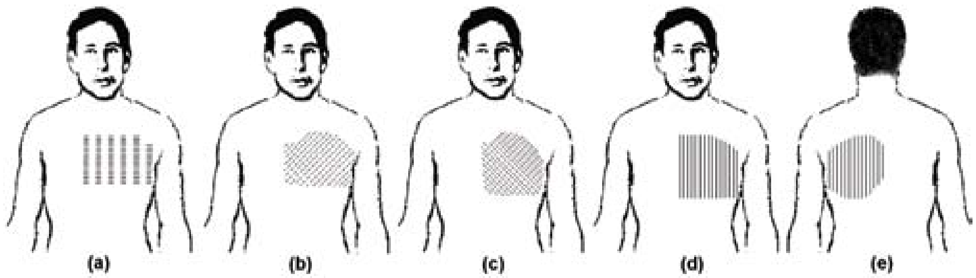
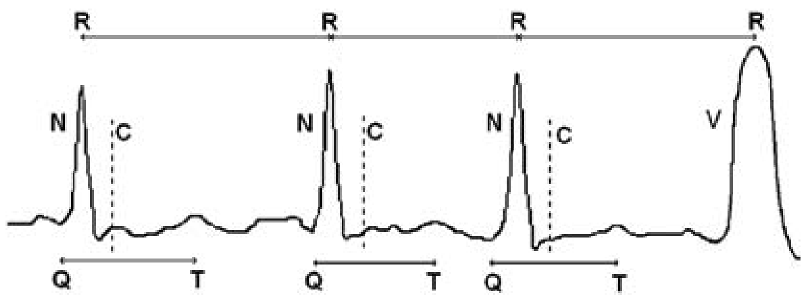


Figure 2. Normal (N) QRS and extra ventricular QRS beat (V), with indicated RR and QT distances. Dotted lines denoted by C (maximal contraction) represent the minimal volume moment of heart during normal cardiac cycle



movement was studied in time and the heart's spatial movement could be reconstructed.

Ultrasound is a particularly harder imaging modality for interpretation than CT and MRI, because it suffers from several specific drawbacks which impede automated analysis. The absence of a relation between pixel intensity and any physical property of the visualized tissue demands a deeper usage of the a priori information. The sensibility to noise and reflections implies an intelligent filtering. Even the definition of border position may be direction dependent (Vuille & Weyman, 1994).

Albeit these afore-named problems, the huge amount of recorded images demands the usage of automatic border detection (ABD) algorithms. We will refrain from a direct comparison of these methods, as there are no standard test data sets neither for this purpose nor standard test criteria (Sher, Revankar, & Rosenthal, 1992). The clinically most widespread algorithm is based on acoustic quantification (AQ) that is involved in several ultrasound machine models. However, this method suffers from high sensitivity to image quality. A typical ABD algorithm contains the following steps:

- Preprocessing (smoothing, contrast checking)
- Edge or region detection (thresholds, edge detectors)
- Geometric object models (for example radial search)
- Anatomical structure model
- Interpretation (based on high-level knowledge)

An excellent example of a structure-based method was published by Geiser and Wilson (Geiser et al., 1998; Sheehan et al., 2000) for automatically detecting endocardial and epicardial borders in short-axis echocardiograms. This method is practically useful and it was validated on a large set of patients. In this method, a great advantage over classic algorithms is the involvement of the a priori information and the usage of a feed-back between a high-level model of a short-axis cross section and a low-level feature data.

Figure 3. Schematic representation of the data recording and analyzing procedure. All echocardiography and ECG data go through the same processing module. The AAM is constructed from the measurements of series 1, and fine tuned afterward using the patient specific data resulting from series 2 stage 1. Stage 2 data serve for the detailed cardiac volumetric analysis. Reconstructed 3-D objects are finally aligned using an iterative LMS-based algorithm.

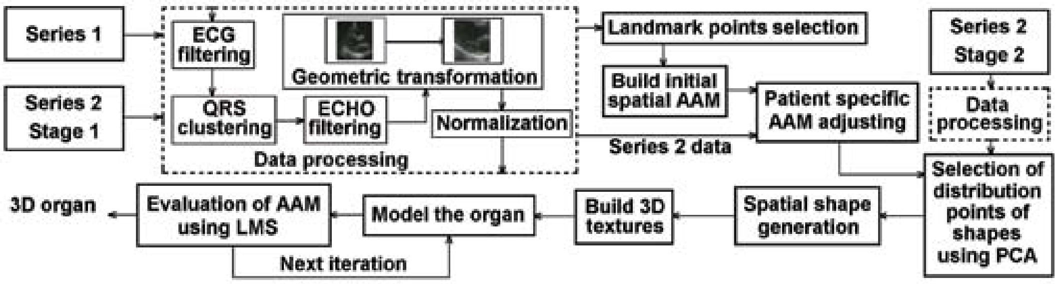
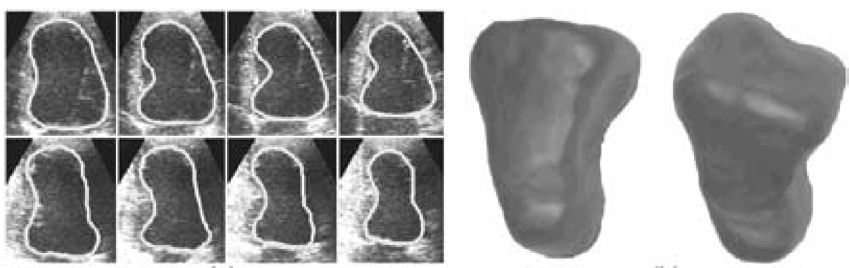


Figure 4. Results of shape reconstruction: (a) Time varying 2-D contour of the left ventricle (2nd patient), (b) reconstructed 3-D structure of the left ventricle



The main weakness lies in its cascade dependent steps that can fail and cause the process to break down.

The enrolled disadvantages convinced us to use an Active Appearance Model that was developed by Edwards, Taylor, and Cootes (1998) for facial recognition and medical image analysis (Edwards et al., 1998) (see Figure 3). This technique can be considered as an extension of the Active Shape Model (ASM) and derives the typical shape and appearance of an echocardiogram from a large set of example images with expert-drawn contours (Bosch et al., 2002; Mitchell et al., 2002).

Its main steps are:

- Extract the average organ shape
- Extract the principal shape variations
- Create the appearance model
- Generate probable echocardiographic image
- Find the desired structure by error minimization technique

This technique has some important advantages: it models average organ shape and variation and can model complete organ appearance including artefact cases. The detailed description of the AAM may be found in an article entitled "Echocardiographic Image Sequence Compression Based On Spatial Active Appearance Model."

RESULTS

Section (a) of Figure 4 presents two series of ultrasound slices indicating the contour of the left ventricle of the second patient, detected during a ventricular contraction. The two rows of slices show two different angle views, having 60° angle difference. The four slices in each row represent subsequent images of the sequence, showing the approximately 100 ms duration of the ventricular contraction. Section (b) of Figure 4 shows two different reconstructed 3-D shapes of the left ventricle, which were obtained using 1527 distribution points.

Figure 5. Variation of heart volume during a normal cardiac cycle (a), and during a ventricular extra cycle (b)

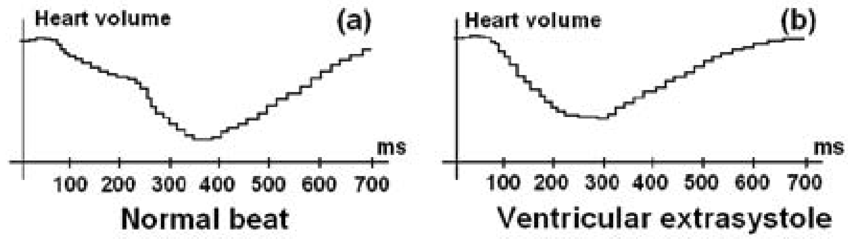
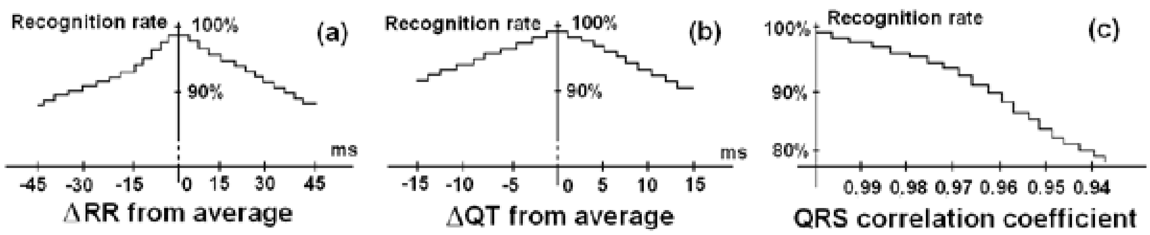


Figure 6. Sensibility of QRS complex recognition algorithm to: (a) Variations of RR distances, (b) variations of QT distances (both with respect to their average), (c) QRS shape correlation with average template.



The evolution in time of the heart volume during a normal and an ectopic (ventricular extra) RR period is presented by Figure 5, sections (a) and (b). As the depolarization wave in case of an ectopic beat starts from the ventricular area, the minimal volume is obtained earlier (the delay caused by the AV-node is skipped), and at a higher value compared to normal beats (the contraction of the heart is not optimally synchronized). The first moment of LV volumetric calculation (t=0 ms) was considered at the moment of maximal positive deflection of the ECG signal.

Figure 6 presents a sensibility analysis of the QRS recognition algorithm. Figure 6, section (a), shows how the variation of the RR distances (shorter or longer than average) influences the recognition rate: the algorithm is more sensible for shorter RR periods. The effect of the variation of QT distances around the average value is shown in section (b) of Figure 6: the same absolute difference makes significantly more damage if it occurs in positive direction. Section (c) of Figure 6 reflects the relation between the shape of QRS complexes and recognition performance. The correlation coefficient of the QRS beat with the cluster averaged shape is directly proportional with the performance of the algorithm.

Table 1 presents the patient-dependent recognition rates for normal beats (average duration 707 ms) and ventricular beats (average duration 671 ms). The measurements were effectuated on four patients, and three of them produced ventricular extra-systolic beats.

The recognition of the relation between echocardiographic image and simultaneously recorded ECG signal is a key element in wall movement detection. This relation can be partially hidden by various events, such as aspiration and expiration. These phenomena may influence the measured heart rate, as presented in Table 2.

The measurement circumstances were selected optimal (average QRS wave shape, average RR and QT distances) for performance values presented in Table 1. As Table 1 reflects, the fourth patient did not produce any ventricular beats. We can observe a pronounced performance advantage of the algorithm for normal beats due its higher incidence.

DISCUSSION

From Figure 6, section (a), we could observe that the variation of the RR distances (shortenings and lengthening) influences the recognition rate in different ways. The relation (connection recognition of the image and electric signal) construction algorithm is more sensible for shorter than average normal RR periods. As Figure

Table 1. Patient-dependent recognition rate of various QRS beats in case of optimal circumstances (average QRS shape, average RR and QT distances)

Patient\ QRS type	Normal QRS beat	Ventricular extra-systole
1st- patient	99.2%	97.6%
2nd- patient	99.4%	92.4%
3rd- patient	98.4%	95.2%
4th- patient	98.6%	Not available data

Table 2. Patient-dependent averaged percentual heart rate modification during aspiration and expiration. The obtained variance is presented for normal and ventricular beats. The fourth patient had no ventricular beats.

Patient\ HR variance	Normal QRS beat	Ventricular extra-systole
1st- patient aspiration	6.2%	6.6%
2nd- patient aspiration	5.5%	3.4%
3rd- patient aspiration	9.2%	7.8%
4th- patient aspiration	5.6%	Not available data
1st- patient expiration	- 4.5%	- 5.7%
2nd- patient expiration	- 5.2%	- 3.4%
3rd- patient expiration	- 6.7%	- 5.9%
4th- patient expiration	- 4.3%	Not available data

6, section (b) presents, the lengthened QT distances deteriorate the recognition performance much more than shortened ones where all QT distances are related to average QT distances for normal beats.

Figure 6, section (c) reflects the relation between the shape of QRS beats and recognition performance. The correlation coefficient of the QRS beat with cluster averaged shape is directly proportional with the algorithm's performance. The identification of a wall segment from an image, measured in time of depolarization (before or in first part of systole) is easier than during repolarization. During a normal RR period, about 21-22 images were recorded, so from such an image sequence the first 7-9 images that are related to depolarization stage could be investigated with higher performance.

The recognition of the relation between echocardiography images and simultaneously recorded ECG signal is a key element in wall movement detection. Various events, such as aspiration and expiration influence the measured heart rate. During a whole cardiac cycle, the shape and volume of the left ventricle changes considerably.

It is difficult to determine the performance of the reconstruction method for the sporadically occurred ventricular extra-systolic beats. Even for patients that produce at least five extra beats with similar shapes in each minute, that are included in the same cluster, the reconstruction performance remains well below the normal QRS cluster's accuracy, due to the sparse distribution of the processable slices.

This relatively smaller population in the cluster of ventricular beats, compared to the normal ones, implies a lower reconstruction performance. The measurement circumstances were selected optimal (average QRS wave shape, average RR, and QT distances) for performance values presented in Table 1. As this table reflects, the fourth patient did not produce any ventricular beats. We can observe a pronounced performance advantage of the algorithm for normal beats due its higher incidence.

The volumetric variance of the heart during normal and ectopic (ventricular extra) RR period is presented by Figure 5, sections (a) and (b). As the depolarization wave in case of ectopic beats starts from ventricular area, the minimal volume is obtained earlier (the AV-node caused delay is skipped) but at a higher value (the contraction of the heart is not optimally synchronized), compared to normal beats.

The visual aspect of the heart and its environment varies in time, due to their mutual motion. AAM models include information concerning inner texture only. It would be beneficial to build another AAM for the modeling of the environment, but the invisibility of the pericardium and lung tissues represents an enormous obstacle.

As we know, the left ventricular (LV) hypertrophy has been identified as a strong and independent predictor of various adverse cardiac events even in patients without associated cardiovascular disease (Benjamin & Levy, 1999). This important clinical parameter has been subject to extensive scientific investigation (Haider, Larson, Genjamin, & Levy, 1998). The conventional 2-D echocardiography is the most commonly used imaging modality that supports LV wall mass determination. However, this method is based on geometric assumptions that do not include some possible abnormal ventricular shapes and variability in the distribution of LV mass, that hardens the development of an accurate, easily reproducible method for the quantification of ventricular mass.

Evaluation of intracardiac anatomy from two dimensional echocardiography image sequences requires a mental conceptualization process that is hardened by cardiac dynamics (MacLeod & Brooks, 1998). The dynamic modeling and visualization systems can help with the spatial interpretation of 3-D data of the heart, and make it possible to build the 3-D model of the heart (Issenberg, Gordon, Gordon, Safford, & Hart, 2001).

This study is an attempt to evaluate whether a virtual modeling and visualizing method is feasible for echocardiography, and if ultrasound images in a virtual reality can advance to a clinically useful tool in the technological process of the future. Unfortunately, the low number of patients and the long processing time reflect that the applicability and benefits of the presented volumetric analyzer and heart modeling method in clinical practice is still limited.

The spatial texture around voxels belonging to a dynamic organ suffers a deformation during contraction. Under such circumstances, a compensation mechanism would be necessary, which is situated beyond the scope of the present article.

CONCLUSION

The investigation of simultaneously recorded ECG and echocardiography images enables us to study the relations between the electrical and mechanical phenomena concerning the heart. The method presented in this article performs correctly in case of normal and ventricular beats, that is, we can monitor the volumetric variance of the heart and its main components. This kind of approach of the problem may result in deeper understanding of several pathological cases like:

- Effects of ectopic beats on the heart's pumping activity
- Risk of development of arrhythmias and fibrillation in case of pathological LV wall thickening
- Mechanical effects of Wolff-Parkinson-White syndrome.

REFERENCES

Angelini, E. D., Laine, A. F., Takuma, S., Holmes, J. W., & Homma, S. (2001). LV volume quantification via spatiotemporal analysis of real-time 3-D echocardiography. *IEEE Trans. Med. Imag., 20*, 457-469.

Belohlavek, M., Manduca, A., Behrenbeck, T., Seward, J., & Greenleaf J. F. (1996). Image analysis using modified self-organizing maps: Automated delineation of the left ventricular cavity in serial echocardiograms. In *Proc. Visual. Biomed. Comp. Conf.* (pp. 247-252).

Benjamin, E. J., & Levy, D. (1999). Why is left ventricular hypertrophy so predictive of morbidity and mortality? *Am. J. Med. Sci., 317*, 168-175.

Bosch, J. G., Mitchell, S. C., Lelieveldt B. P. F., Nijland, F., Kamp, O., Sonka, M., et al. (2002). Automatic segmentation of echocardiographic sequences by active appearance motion models. *IEEE Trans. Med. Imag., 21*, 1374-1383.

Brotherton, T., Pollard, T., Simpson, P., & DeMaria, A. (1994). Echocardiogram structure and tissue classification using hierarchical fuzzy neural networks. In *Proc. IEEE Conf. Acoustics Speech Sign. Proc.,* (Vol. 2, pp. 573-576).

Chalana, V., Linker, D. T., Haynor, D. R., & Kim Y. (1996). A multiple active contour model for cardiac boundary detection on echocardiographic sequences. *IEEE Trans. Med. Imag., 15*, 290-298.

Cootes, T. F., Edwards, G. J., & Taylor, C. J (2001).

Active appearance models. *IEEE Trans. Patt. Anal. Mach. Intell, 23*, 681-685.

Dias, J. M. B., & Leitão, J. M. N. (1996). Wall position and thickness estimation from sequences of echocardiographic images. *IEEE Trans. Med. Imag., 15,* 25-38.

Duncan, J. S., & Ayache, N. (2000). Medical image analysis: Progress over two decades and the challenges ahead. *IEEE Trans. Patt. Anal. Mach. Intell., 22*, 85-105.

Edwards, G. J., Taylor, C. J., & Cootes, T. F. (1998). Interpreting face images using active appearance models. In *Proc. 3rd Int. Conf. Face and Gesture Recogn,* Japan (p. 300).

Geiser, E. A., Wilson, D. C., Wang, D. X., Conetta, D. A., Murphy, J. D., & Hutson, A. D. (1998). Autonomous epicardial and endocardial boundary detection in echocardiographic shor-axis images. *J. Amer. Soc. Echocardiogr., 11*, 338-348.

Haider, A. W., Larson, M. G., Benjamin, E. J., & Levy, D. (1998). Increased left ventricular mass and hypertrophy are associated with increased risk for sudden death. *Journ. Am. Coll. Cardiol., 32,* 1454-1459.

Issenberg, S. B., Gordon, M. S., Gordon, D. L., Safford, R. E., & Hart, I. R. (2001). Simulation and new learning technologies. *Med. Teach, 23,* 16-23.

Lagerholm, M., Peterson, C., Braccini, G., Edenbrandt, L., & Sörnmo, L. (2000). Clustering ECG complexes using hermite functions and self-organizing maps. *IEEE Trans. Biomed. Eng., 47*, 838-848.

MacLeod, R. S., & Brooks, D. H. (1998). Recent progress in inverse problems in echocardiology. *IEEE EMBS Mag., 17(5),* 73-83.

Mitchell, S. C., Bosch, J. G., Lelieveldt, B. P. F., van der Geest, R. J., Reiber, J. H. C., & Sonka, M. (2002). 3-D active appearance models: segmentation of cardiac MR and ultrasound images. *IEEE Trans. Med. Imag., 21*, 1167-1178.

Mitchell, S. C., Lelieveldt, B. P. F., Bosch, J. G., van der Geest, R. J., Reiber, J. H. C., & Sonka, M. (2002). Segmentation of cardiac MR volume data using 3D active appearance models. In *Proc. SPIE Med. Imag*. (pp. 433-443).

Montagnat, J., Delingette, H., & Malandain, G. (1999). Cylindrical echocardiographic image segmentation based on 3-D deformable models. *Lect. Notes Comp. Sci. (MICCAI' 99), 1679,* 168-176.

Sheehan, F., Wilson, D. C., Shavelle, D., & Geiser E. A. (2000). Echocardiography. In M. Sonka, & J. M. Fitzpatrick (Eds.), *Handbook of medical imaging. Vol 2: Medical image processing and analysis* (pp. 609-674). Bellingham, WA: SPIE.

Sher, D. B., Revankar, S., & Rosenthal, S. (1992). Computer methods in quantitation of cardiac wall parameters from two dimensional echocardiograms: A survey. *Int. J. Card. Imaging*, 8-11.

Stegmann, M., & Pedersen, D. (2005). Bi-temporal 3D active appearance models with applications to unsupervised ejection fraction estimation. *Progr. Biomed. Opt. Imag. Proc., SPIE 5746,* 336-350.

Szilágyi, S. M., Benyó, Z., & Dávid, L. (2003). Iterative ECG filtering for better malfunction recognition and diagnosis. In *Proc. 5th Int. IFAC Symp. Modell. Contr. Biomed. Syst., Melbourne* (pp. 295-300).

Szilágyi, S. M., Benyó, Z., & Szilágyi, L. (2005). Recognition of various events from 3–D heart model. In *16th IFAC World Congress Preprint,* Prague (Paper #4695, pp. 1-6).

Vuille, C., & Weyman, A. E. (1994). Left ventricle I: General considerations, assessment of chamber size and function. In A. E. Weyman (Ed.), *Principles and practice of echocardiography* (p. 56). Philadelphia, PA: Lea & Febiger.

KEY TERMS

Active Appearance Model (AAM): A computer vision algorithm for matching a statistical model of object shape and appearance to a new image.

Active Shape Model (ASM): The statistical model of the shape of objects which iteratively deforms to fit to an example of the object in a new image.

Anatomical Structure Model: An anatomical model of the studied organ.

ECG Analysis: The collection of ECG signal processing methods aiming at the determination of medical parameters.

Echocardiogram: The image of the heart created with ultrasound imaging technology.

QRS Complex: A structure on the ECG signal that corresponds to the depolarization of the ventricles.

Spatio-Temporal Modeling: Modeling in both time and space.

Volumetric Analysis: The analysis of the volume variance in time.

Weakness of Modern Hospitals

Roberto Setola
Università Campus Bio-Medico di Roma, Italy

INTRODUCTION

Until some decades ago, health care services were primarily supplied inside hospitals. The patient had to move from his or her home to the hospital, where various diagnostic and therapeutic treatments were provided. Moreover, inside the same hospital, the various tools and processes were insulated and autonomous. Patients and doctors had to move from one tool to another, often placed in different areas, to acquire the different resources. Information provided by these tools was generally collected via paper records. These records were then physically moved (generally in a very inefficient way) from one place to another in order to exchange information. In many situations, these records represented the most critical element in the health system due to misunderstanding, errors, and loss of information caused by their use. Nevertheless, its main drawback was the difficulty, or even the impossibility, to retrieve information from paper records when there were needs in the future and/or outside the hospital.

This situation was largely inefficient, especially from the patient's point of view. Indeed, he or she had to spend a lot of time moving to and from the hospital and supporting a great deal of stress to supply time and again the same information (often in an incomplete, vague, or erroneous manner), and eventually suffering for errors and/or unavailability of information previously stored in other records.

To overcome this drawback, in the past few years, the health care systems has largely changed the way in which their services are produced and supplied. In order to improve the quality (e.g., better diagnosis and treatments), guarantee more comfort to the patients, increase the efficiency of the systems (e.g., reducing time in hospital and, more generally, any type of side-effects), and have a more rational use of the money, ICT (information and communication technologies) has spread throughout fields of health care. Indeed, due to the power of ICT to make interoperable the different elements and actors, useful synergies have been exploited to better use any resource.

This contributes to improve the quality of the services (qualified and specialized services may be provided in spite of geographical location) and, at the same time, to improve efficiency (because it is possible to better distribute loads and to exploit synergy and scale economy).

To achieve this result, the various apparatuses, components, and systems (once autonomous entities) are no longer insulated but rather are grouped into networks to allow sharing information, to support cooperation, to exploit complementary, and to implement supplementary strategies.

In this framework, from the patient's point of view, any single element of the system is less and less important owing to the network as a whole providing the services. This is obtained by integrating multichannels delivery strategies into the front-ends and making interoperable the various back-end elements.

Implicitly in this change of paradigm is the increased rule played by technological infrastructures in the health care systems. These infrastructures that, until few years ago, were not strictly related with the core business (i.e., supply care) but rather confined to complementary activities, today are becoming the backbone of any health care system.

Unfortunately, this introduces many *dependency* and *interdependency* links among the various components. This represents the real weakness of this scenario. Indeed, even if a network-based health care system is more robust than a model composed by single "asset" with respect to a component's failure, it appears more fragile to "catastrophic" events. The presence of these interdependencies (many of them not designed or considered at implementation time and actually poorly known or even completely hidden) exposes the system to a huge variety of threats. Moreover, it makes the system susceptible, due to domino and cascade effects, to simultaneous failures of many services, as dramatically emphasized in health care and other sectors by the blackouts of 2003 (Italian Government Working Group, 2004).

Then, moving from a traditional health care scenario to a network-centered framework, we improve our capabilities to provide efficient, effective, and economic services, but at the same time, we also need to consider the side-effects so introduced. This chapter is devoted specifically to pinning down some of them and to stimulating great attention to this topic.

BACKGROUND

To analyze the amplification of negative consequences of a failure owing to the presence of interdependencies among the various infrastructures used inside a modern hospital, we have used the Input-output Inoperability Model (IIM). IIM is a simple tool, proposed by Haimes and Jiang (2001), to emphasize how the presence of dependencies and interdependencies among the various components of a complex system may facilitate the spreading of degradation.

Haimes and Jiang (2001) set up this model, building on the well-known theory on market equilibrium by Nobel Prize winner Wassily Leontief. The IIM uses the same framework proposed by Leontief but instead considers how the production of goods or services of a firm influences the level of production of the other firms; it focuses its attention on the spreading of "degradation" into a networked system. To this end, the authors introduce the concept of *inoperability*, defined as the inability of a system to perform its intended functions and analyze how a given amount of inoperability inside one element influences the other components of the network.

Haimes, Horowotz, Lambert, Santos, Lian, and Crowther (2005) use this approach to analyze how inoperability induced by a High Altitude Electromagnetic Pulse (HEMP) affects the various sectors of the U.S. economy and to estimate the recovery time under various hypotheses. In Reed, Chang, and McDaniels (2006), this approach has been used to provide a tool for resource allocation for postevent recovery considering the Katrina hurricane scenario, while in Panzieri and Setola (2008), the approach is modified to also explicitly consider the spreading of failures.

The great interest in this approach is related to its simplicity, even if the results that it provides are, for many aspects, largely qualitatively and oversimplified.

The idea at the base of IIM is that an event (e.g., a failure) that reduces the capability of the *i*-th infrastructure to correctly operate induces degradation also in other infrastructures that use services or goods produced by the *i*-th one. This degradation may further propagate involving other infrastructures (cascade phenomena) or even exacerbate the negative consequences into the *i*-th one (feedback effect).

Mathematically, IIM describes these phenomena on the basis of the level of inoperability associated with the various infrastructures. Specifically, the inoperability of the *i*-th infrastructure is coded via the variable x_i defined in the range [0, 1], where $x_i = 0$ means that the infrastructure is fully operative, while $x_i = 1$ means that the infrastructure is completely inoperable.

The inoperability induced on the system due to persistent external causes u_i is calculated via the following dynamic equation:

$$\mathbf{X}(k+1) = \max\left\{\mathbf{A}\,\mathbf{X}(k) + \mathbf{U}, \mathbf{1}\right\} \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^n$ and $\mathbf{U} \in \mathbb{R}^n$ are vectors composed, respectively, by the level of inoperability and external failure associated with each one of the *n* infrastructures considered in the scenario. $\mathbf{A} \in \mathbb{R}^{n\times n}$ is the Leontief matrix, which entry a_{ij} represents the level of influence that the inoperability of the *j*-th infrastructure has on the *i*-th one.

We impose that $a_{ii} = 0 \;\; \forall i$ because we do not consider any recovery phenomena. Notice that in the model, $a_{ij} = 1$ means that the *i*-th infrastructure is completely dependent on the *j*-th one, because a given amount of failure in the latter will directly induce an equal level of degradation into the *i*-th infrastructure.

In order to evaluate the level of dependencies of an infrastructure, we introduce the *dependency index*, defined as the sum of the Leontief coefficient along a single row:

$$\delta_i = \sum_j a_{ij} \quad \text{(row summation)} \tag{2}$$

This index represents a measurement of the robustness of the corresponding infrastructure with respect to the inoperability of the others. If this quantity is less than 1, then the *i*-th infrastructure preserves some working capabilities (e.g., thanks to the presence of buffers, UPS, etc.) in spite of the level of inoperability of its suppliers. On the other side, when $\delta_i > 1$, the operability of the *i*-th infrastructure may be completely nullified, even if the supplier infrastructures show some residual capabilities.

The influence that a specific infrastructure has on the global system may be evaluated, considering the *influence gain* (i.e., the sum of the Leontief coefficients along a single column)

$$\rho_j = \sum_i a_{ij} \quad \text{(column summation)} \qquad (3)$$

A large value for this index means that the inoperability of *j*-th infrastructure will induce significant degradation into the whole system. When $\rho_j > 1$, the negative effects induced by cascade phenomena on the other infrastructures are more relevant in terms of inoperability than that affecting the *j*-th one. The opposite happens when $\rho_j < 1$.

FRAGILITY OF A MODERN HOSPITAL

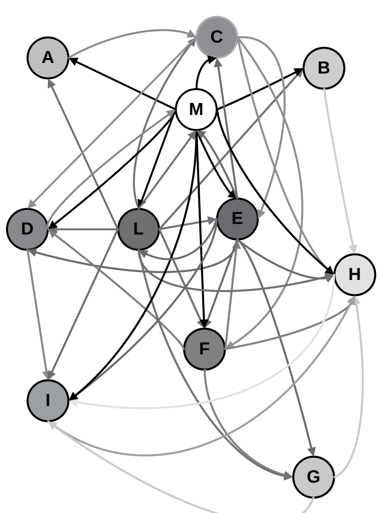
We use IIM to analyze the impact of a failure in the *IP-network* of a hospital comparing the consequences induced in a modern hospital with that registered in a traditional one.

Specifically, we applied this model to the actual and to the in-building hospital of our university. Indeed, actually the CAMPUS has a traditional hospital, but it is building a new one with a very strong modern idea.

We adopt a very crude approximation modelling each technological infrastructure presents inside the hospital as an entity that the level of operability depends, further to external causes, on the availability of "resources" supplied by other infrastructures. The technological infrastructures on which we focus our analysis are reported on the right of Figure 1.

In a modern hospital, as reported in Figure 1, technological infrastructures figure out a very complicated web of reciprocal influences. Inside this *Electric network* (L), *Infrastructure monitoring* (M) and *IP network* (E) represent hubs with the highest out-degree (i.e., number of outgoing links that are 10, 9, and 8, respectively). This means that these infrastructures have a large influence on the others (they have many non-null terms in the influence gain). On the other side, *Bio-medical apparatus monitoring* (H) has the highest in-degree (i.e., number of ingoing links that is equal to 9). Then the capability of this infrastructure to operate correctly largely depends on the operability of the other infrastructures. Consequently, in the corresponding row of the Leontief matrix, as shown in Table 1, many elements are different from zero, which suggests that the corresponding *dependency index* is quite large, making more "concrete" the risk that this

Figure 1. Inoperability influence graph for the modern hospital



A – water

B – medical gas

C – air conditioning

D – telephone

E – IP network

F – IT infrastructure

G – (electronic) case history

H – biomedical apparatus monitoring

I – medical alarms

L – electric network

M – infrastructures monitoring

Table 1. Leontief matrix for a traditional hospital. Notice that in this case, (G) represents paper-based case history. The matrix has been bordered with the dependency δ_i and influence ρ_j indexes.

	A	B	C	D	E	F	G	H	I	L	M	$\delta_i = \sum_j a_{ij}$
A	0	0	0	0	0	0	0	0	0	0	0,03	0,03
B	0	0	0	0	0	0	0	0	0	0,3	0,1	0,40
C	0,03	0	0	0	0	0	0	0	0	0,6	0,05	0,68
D	0	0	0,10	0	0	0,01	0	0	0	0,15	0,03	0,29
E	0	0	0,15	0	0	0,10	0	0	0	0,3	0,05	0,60
F	0	0	0,20	0	0,25	0	0	0	0	0,2	0,05	0,70
G	0	0	0	0	0	0	0	0	0	0,05	0	0,05
H	0,05	0,15	0,20	0	0,10	0,05	0	0	0,05	0,3	0,05	0,95
I	0	0	0	0,10	0,05	0	0	0	0	0,25	0,20	0,60
L	0	0	0	0	0	0	0	0	0	0	0,07	0,07
M	0	0	0	0	0,03	0	0	0	0	0,08	0	0,09
$\rho_j = \sum_i a_{ij}$	0,08	0,15	0,65	0,10	0,43	0,16	0	0	0,05	2,23	0,63	

Table 2. Leontief matrix for a modern hospital. Notice that in this case, (G) represents electronic case history. The matrix has been bordered with the dependency δ_i and influence ρ_j indexes.

	A	B	C	D	E	F	G	H	I	L	M	$\delta_i = \sum_j a_{ij}$
A	0	0	0	0	0	0	0	0	0	0,05	0,06	0,11
B	0	0	0	0	0	0	0	0	0	0,30	0,13	0,43
C	0,03	0	0	0	0,08	0	0	0	0	0,60	0,08	0,79
D	0	0	0,10	0	0,40	0,10	0	0	0	0,15	0,04	0,79
E	0	0	0,15	0	0	0,18	0	0	0	0,35	0,09	0,77
F	0	0	0,20	0	0,43	0	0	0	0	0,23	0,06	0,92
G	0	0	0	0	0,63	0,87	0	0	0	0,39	0	1,89
H	0,05	0,15	0,20	0	0,24	0,19	0,31	0	0,05	0,42	0,08	1,69
I	0	0	0	0,08	0,09	0	0,02	0,02	0	0,29	0,21	0,71
L	0	0	0	0	0,02	0	0	0	0	0	0,09	0,11
M	0	0	0	0,14	0,04	0	0	0	0	0,22	0	0,40
$\rho_j = \sum_i a_{ij}$	0,08	0,15	0,65	0,22	1,93	1,34	0,32	0,02	0,05	3,00	0,84	

infrastructure exhibits large degradations owing to failures in other infrastructures.

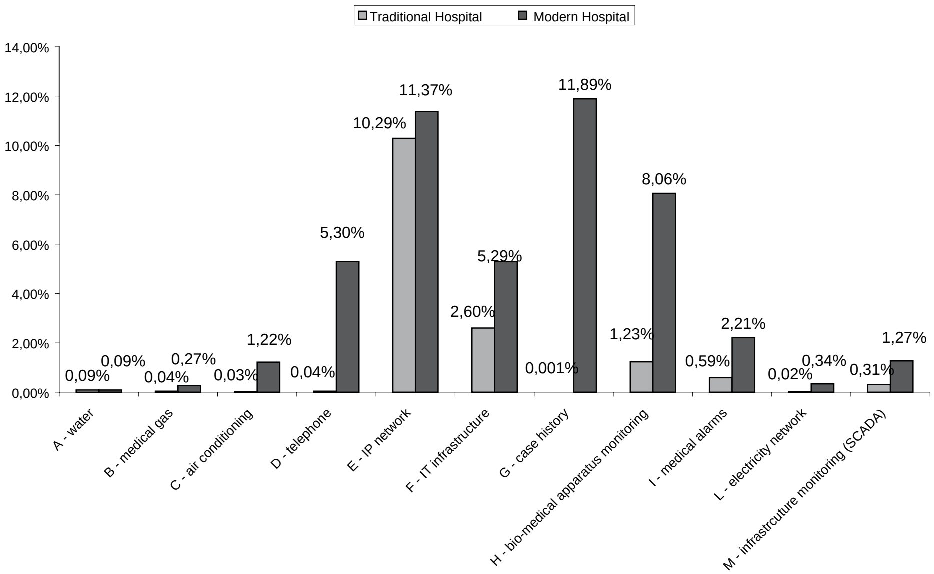
However, to better understand the role played by the various infrastructures and to identify the most critical elements, we have to also consider the degree of influence that characterizes the various links.

This represents the hardest task to be performed in order to apply IIM. In Haimes, et al. (2005), the Leontief matrix **A** is evaluated using the economical statistical data provided by the Bureau of Economic Analysis (BEA). In this chapter, we follow a different strategy evaluating these coefficients on the basis of information collected directly on the field.

To this end, we interviewed managers, architects, engineers, and technicians involved in the design of the various infrastructures for the new hospital. This information has been merged with "experiences" coming from doctors and medical assistants. Moreover, as proposed in Panzieri and Setola (2008), we also used schemas and maps of the infrastructures to also evaluate physical, geographical, and cyber interdependencies (for more information on this taxonomy, see Rinaldi, Peerenboom, and Kelly, 2001).

W

Figure 2. Impact induced by a failure that reduces 10% of the operability of the IP-network (E) in a traditional and a modern hospital. Notice that in the traditional hospital, the case history (G) is paper-based, while in the modern hospital, it is electronic.



Comparing the Leontief matrices of Tables 1 and 2, one can notice that in the modern hospital, there are more no-null elements, and that except for the relevance of the *telephone network (D)*, there is a generalized augment of the influence coefficients.

In Table 1, the *dependency index* δ_i varies from 0,05 (associated with the paper-based *Case history* (G)) to 0,95 (of the *Biomedical apparatus monitoring* (H)). Because this quantity is considerably less than 1 for quite all the infrastructures, we may infer that in a traditional hospital, the various infrastructures were substantially autonomous and able to correctly supply their services without the need of resources supplied by the other infrastructures.

On the other side, in a modern hospital (Table 2), this quantity varies from 0,11, corresponding to *Water infrastructure* (A), to 1,89 related with the *Electronic case history* (G). Moreover, *Biomedical apparatus monitoring* (H) shows a very high value (1,69). This means that there is a generalized interdependencies increment but also an augment of fragility because the failure of any infrastructure has a direct influence on the capability of the other infrastructures to correctly perform their own works.

Curiously, the case history infrastructure that in the traditional hospital is one of the less sensitive elements with respect to infrastructures' failures, becomes in the modern hospital (where the electronic version has been adopted) one of the most sensitive ones.

The level of influence varies for the traditional hospital, from 0 (*Case history* (G) and *Biomedical apparatus monitoring* (H) have no impact on the other infrastructures) to 2,23 of the *Electric network* (L). The large value of this last coefficient emphasizes the cornerstone role played by electricity in the traditional hospital.

If we look to the modern hospital, we see that the influence gain varies from 0,02 (associated with the *Bi-medical apparatus monitoring* (H)) to the value 3 of the *Electric network* (L). However, in this scenario, the influence gain associated with the *IP-network* (E) and that of the *IT-infrastructure* (F) are 1,93 and 1,34, respectively. This pins down two elements. First, the importance of the electric network is increased, and it is still the most critical infrastructure within the hospital. Second, as mentioned before, the capability of each infrastructure to autonomously operate has been reduced due to the need to use services provided by the others. Obviously, this makes the whole system more prone to amplify negative consequences owing cascade phenomena.

To compare the behavior of the two configurations, we analyze the overall consequences induced by a persistent failure that reduces 10% of the operability of the *IP network* (E) evaluating the steady-state solution of equation (1). The results are reported in Figure 2.

Looking to the traditional hospital, one observes that the other infrastructures are only marginally influenced by the failure in the *IP-network* (E). Only *IT Infrastructure* (F) and *Biomedical apparatus monitoring* (H) show a degradation greater than 1%. Moreover, due to the presence of feedbacks, the inoperability of the *IP network (E)* is slightly increased.

On the other side, the consequences of the same failure in the modern hospital are more relevant. It is evident that in this case there is considerable increment in the *IP-network (E)* inoperability (due to feedback) but also a generalized diffusion of inoperability. Indeed, except for *Water* (A), *Medical Gas* (B), and *Electricity* (L), all the infrastructures show an inoperability level greater than 1%. The highest rate of inoperability is shown by the *Electronic case history* infrastructure (G). It reaches a degradation of about 12%.

Looking to Table 3, one can recognize that the direct influence of the *IP-network* (E) on the *Electronic case history* (G) is $a_{GE} = 0,63$. Then, neglecting interdependency phenomena, we should have predicted a degradation of about 6%. Nevertheless, as shown, the presence of the interdependencies amplifies the consequences of the failure, producing an inoperability level that is quite the double of that foreseen in an atomistic analysis.

Obviously, modern infrastructures are designed, conducted, and managed with higher standards in order to improve their efficiency but also their robustness and business continuity. Then it is very difficult that the same "event" may cause similar consequences in both scenarios. Indeed, in everyday activities, we observe that in modern hospitals, quite all the consequences of negative events that affect, as in our example, IP-networks are absorbed without any appreciable degradation (thanks to redundancy, backup elements, etc.). However, as our analysis and different real-world episodes have shown, if an event is able to induce tangible degradation into a single infrastructure, it is largely amplified due to interdependencies and domino effects to a point that it may induce generalized inoperability in the whole system.

FUTURE TRENDS

It is evident that the in the next years, any system, including hospitals and health care systems, will use more and more ICT and other technological infrastructures. This will increment their networked-centered profile and, consequently, the presence of a huge number of dependencies and interdependencies within any element. This will contribute to improve the capability to provide very efficient services, but at the same time, as emphasized, it will make the system more prone to very catastrophic failures.

Then, it is mandatory to correctly figure out this complex scenario in order to understand the possible side-effects induced by the spread of ICT. Unfortunately, this represents a very difficult task, because actual methodological tools appear largely unaffected due to the complexity that characterizes interdependent frameworks. However, we should quickly learn how to approach this scenario in order to adequately manage its behavior.

This will represent one of the most important challenges for the next millennium.

CONCLUSION

An emergent paradigm in developed countries is the increased relevance of networked systems in any sectors. Indeed, technological, economical, sociological, and political reasons suggest, or better impose, to adopt a service-oriented paradigm where the customer (the patient) is collocated in the center of a complex web of technological infrastructures.

Specifically, in the health care framework, exploiting the facilities provided by ICT, it is possible to supply customized health services with very high quality, and better comfort for the patient; at the same time, we are able to reduce their costs.

However, this scenario, mainly dominated by the interdependencies phenomena, has to be carefully understood in order to manage its side-effects. Indeed, even if this framework appears very robust with respect to "normal" accidents, it is prone to some "rare" events (actually neither identifiable nor predictable) that may induce catastrophic and generalized failure in the whole system.

W

REFERENCES

Haimes, Y. Horowotz, B., Lambert, J., Santos, J., Lian, C., & Crowther, K. (2005). Inoperability input-output model (IIM) for interdependent infrastructure sectors: Theory and methodology. *Journal of Infrastructure Systems*, 67–79.

Haimes, Y., & Jiang, P. (2001). Leontief-based model of risk in complex interconnected infrastructures. *Journal of Infrastructure Systems*, 1–12.

Italian Government Working Group on Critical Information Infrastructure Protection. (2004). *La protezione delle infrastrutture critiche informatizzate—La realtà Italiana* (in Italian).

Panzieri, S., & Setola, R. (2008). Failures in critical interdependent infrastructures. *Int. J. Modelling, Identification and Control (IJMIC)*.

Reed, D., Chang, S., & McDaniels, T. (2006). *Modelling of infrastructure interdependencies*. Proceedings of the Third International Conference on Critical Infrastructure (CRIS), Alexandria, Virginia.

Rinaldi, S., Peerenboom, J., & Kelly, T. (2001). Identifying understanding and analyzing critical infrastructure interdependencies. *IEEE Control System Magazine*, pp. 11–25.

KEY TERMS

Asset: Element able to supply a specific service to an easily identifiable set of users (generally located in the proximity of the asset itself).

Dependency: The capability of an infrastructure to influence the state of another infrastructure; a *unidirectional* relationship.

Infrastructure: System composed by many interoperable elements, geographically dispersed, able to supply generalist products (that have to be specialized by the end-user) to an unforeseen set of users, difficult to identify and to estimate.

Inoperability: The inability of a system to perform its intended functions.

Interdependency: A *bidirectional* relationship between two infrastructures through which the state of each infrastructure is influenced or is correlated to the state of the other.

Service: Any activity identifiable by a tangible benefit for the end-user.

Wearable Kinesthetic System for Joint Knee Flexion-Extension Monitoring in Gait Analysis

Mario Tesconi
University of Pisa, Italy

Enzo Pasquale Scilingo
University of Pisa, Italy

Pierluigi Barba
University of Pisa, Italy

Danilo De Rossi
University of Pisa, Italy

INTRODUCTION

Posture and motion of body segments are the result of a mutual interaction of several physiological systems such as nervous, muscle-skeletal, and sensorial. Patients who suffer from neuromuscular diseases have great difficulties in moving and walking, therefore motion or gait analysis are widely considered matter of investigation by the clinicians for diagnostic purposes. By means of specific *performance tests*, it could be possible to identify the severity of a neuromuscular pathology and outline possible rehabilitation planes. *The main challenge is to quantify a motion anomaly, rather than to identify it during the test.* At first, visual inspection of a video showing motion or walking activity is the simplest mode of examining movement ability in the clinical environment. It allows us to collect qualitative and bidimensional data, but it does not provide neither quantitative information about motion performance modalities (for instance about dynamics and muscle activity), nor about its changes. Moreover, the interpretation of recorded motion pattern is demanded to medical personnel who make a diagnosis on the basis of subjective experience and expertise. A considerable improvement in this analysis is given by a technical contribution to quantitatively analyse body posture and gesture. Advanced technologies allow us to investigate on anatomic segments from biomechanics and kinematics point of view, providing a wide set of quantitative variables to be used in multi-factorial motion analysis. A personal computer enables a real-time 3D reconstruction of motion and digitalizes data for storage and off-line elaboration. For this reason, the clinicians have a detailed description of the patient status and they can choose a specific rehabilitation path and verify the subject progress.

In this context, the Gait Analysis has grown up and currently provided a rich library about walking ability.

In a typical Motion Analysis Lab (MAL), multiple devices work together during walking performance, focusing on different motion aspects, such as kinematics, dynamics, and muscular activity.

The core of the equipment used in the MAL is a stereophotogrammetric system that measures 3D coordinates of reflecting markers placed on specific anatomic "repere" on the body (Capozzo, Della Croce, Leardini, & Chiari, 2005). This instrumentation permits to compute angles, velocities, and accelerations.

The MAL is also equipped with force platforms, which reveal force profiles exchanged with ground on landing.

The electromyographic unit is deputed to measure muscle contraction activity by using surface electrodes. Pressure distributions can be obtained by means of baropodometric platforms, made up of a matrix of appropriately shaped sensors. A unique software interface helps the operator to simply manage data from the complex equipment.

BACKGROUND

Although its remarkable advantages, the Quantitative Gait Analysis require large spaces, appropriate environments, and cumbersome and expensive equipment that limit the use to restricted applications.

Moreover, the stereophotogrammetric system requires a pretest calibration and complex procedure which consists in the placement of reflecting markers on the subject body.

Electromyography may be obtrusive if needle electrodes are used to investigate deeper muscles or single motor unit activity. Because of this, although it is a widely accepted methodology, Gait Analysis is still quite difficult to be used in clinical contexts.

For these motivations, in the last few years technological research accepted the challenge of transferring any support and high level information close to the patient, better on the patient himself, by creating a *user friendly* patient-device interface. This means that health services can be brought out of their classic places with a remarkable improvement in health service and in health care based on prevention. Patient can be monitorized for longer time, ensuring the same high service quality at lower costs. A profitable combination of expertise and know how from electronic engineering, materials science and textile industry led to develop effective *Wearable Health Care Systems* (or simply *Wearable Systems*) (Lymberis & De Rossi, 2004). These latter ones had electronics integrated on traditional clothes, like on a breadboard. In order to allow subjects to freely move without constraints, improvements in materials manufacture, microelectronics, and nanotechnologies allowed us to develop a new generation of wearable sensing systems characterized by multifunctional fabrics (referred to as *e-textiles* or *smart textiles*), where electronic components are made up of polymeric materials, weaved or impressed on the fabric itself, without losing their original mechanical and electrical properties.

The main advantages of this advanced technology consist of:

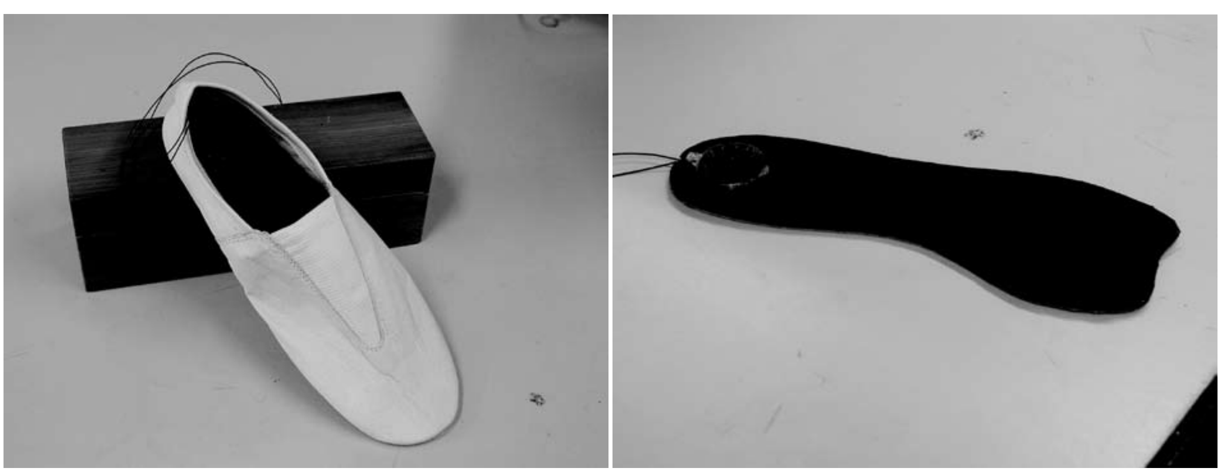
- Strict contact of fabric with skin, the natural human body interface
- Flexibility and good adherence to the body, which helps avoiding motion artefacts affecting the measurement
- Washability to be reused
- Lightness and unobtrusivity, so the patient can naturally move in his usual environment

By means of wearable systems, a large set of physiological variables (ECG, blood pressure, cardiac frequency, etc.) can constantly be monitored and quantitative parameters can be computed, from motion sensing, by dedicated software. Important applications related to health care deal with Ergonomics, Telemedicine, Rehabilitation, and any health service (elderly or impaired people).

MATERIALS AND METHODS

As described in the introductive section, Gait Analysis is usually performed for measuring kinematic variables

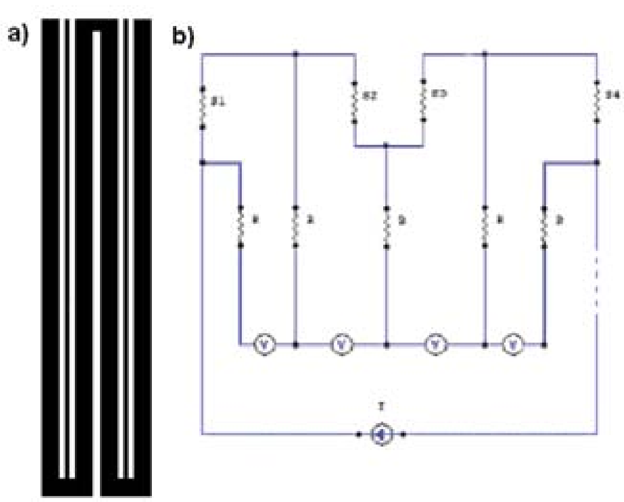
Figure 1. Sensorized shoe for marking the gait cycle



of anatomic segments with tracking devices, which have their main disadvantages in being complex, expensive, and not easy to be used in clinical environments. Thanks to the development of wearable systems, a new class of smart textiles (De Rossi, Santa, Mazzoldi, 1999) allows us to design and produce special garments using electrically conductive elastomer composites (CEs) (Post, Orth, Russo, & Gershenfeld, 2000). CEs show piezoresistive properties when a stress is applied, and in several applications they can be integrated into fabrics or other flexible substrates acting as strain sensors (Scilingo, Lorussi, Mazzoldi, & De Rossi, 2003). Such a sensors are just applied in sportive medicine for rehabilitation in soccer post-strokes, for athletic gesture analysis, for angles measurement, and for lower limb force analysis. We used a wearable prototype based on CEs sensors for the first time in a clinical environment. It consists of a sensorized fabric bend wrapped around the knee, very light, well adherent and elastic, hence very suitable for the chosen application. A sensing garment has been realized deposing the CEs over a sublayer of cotton Lycra® and building a sensorized shoe devoted to the gait cycle synchronism. The sensorized shoe consists of a pressure sensor realized of the same material of the knee-band (KB), and is placed on the sole near the heel, as shown in Figure 1.

The CEs mixture is smeared on the fabric previously covered by an adhesive mask cut by a laser milling machine. The mask is designed according to the shape and the dimension desired for sensors and wires. After smearing the solution, the mask is removed. Then the treated fabric is placed in an oven at a temperature of about 130 centigrade degrees. During this phase, the cross-linking of the solution speeds up, and in about 10 minutes the sensing fabric is ready to be employed. This methodology is used both for sensing areas and connection paths, thus avoiding employing metallic wires to interconnect sensors or link them to the electronic acquisition unit (Lorussi, Rocchia, Scilingo, Tognetti, & De Rossi, 2004). In this way, no rigid constraints or linkages are present and movements are unbounded (Tognetti et al., 2005).

Figure 2. (a) The mask used for the sensorised KB. (b) The equivalent electric scheme of the KB.

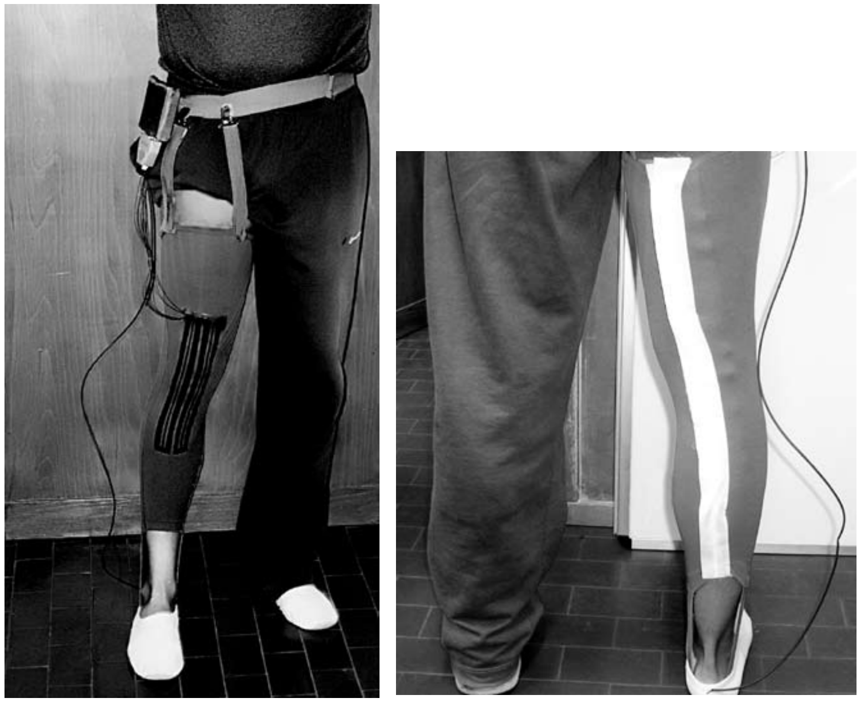


Section (a) of Figure 2 shows the mask for the realization of the KB; the bold black track represents four sensors connected in series and covers the knee joint. The thin tracks represent the connection between the sensors and the electronic acquisition system. Being the thin tracks are made of the same piezorestive mixture, they undergo a not negligible (and unknown) change in their resistance when the knee bends. To face this inconvenience, the analog front-end of the electronic unit is designed to compensate the resistance variation of the thin tracks during the deformations of the fabric. The electric scheme is shown in section (b) of Figure 2. While a generator supplies the series of sensors S_i with a constant current I, the acquisition system is provided by a high input impedance stage realized by instrumentation amplifiers and represented in section (b) of Figure 2 by the set of voltmeters. The voltages measured by the instrumentation amplifiers are equal to the voltages which fall on the S_i that is related to resistances of the sensors. In this way, the thin tracks perfectly substitute the traditional metallic wires and a sensor consisting of a segment of the bold track between two thin tracks can be smeared in any position to detect the movements of a joint. The KB acquires information on the flexion-extension of the knee from four sensors spread on a leotard and a step-signal from the sensorized shoe (Morris & Paradiso, 2002). Moreover, the prototype consisted of an "on body unit" dedicated to the acquisition of signals from the KB and the Bluetooth transmission to a remote PC as shown in Figure 3.

EXPERIMENTAL PROTOCOL

The prototype was used in a clinical application consisting of the gait monitoring on subjects affected by

Figure 3. The overall prototype consisting of a KB, sensorized shoe and an "on body unit." Front (left side) and back (right side) views



venous ulcers localized in the lower limbs. A test for discriminating between pathologic and normal walking was performed by the acquisition of flexion-extension signals from the knee joint during movement. Next, experimental data were validated in order to assess goodness of the methodology. Nine subjects, five females and four males, volunteered to participate in the study. The "normal" sample was made of four voluntaries, three males and one female, belonging to university context with a mean age of 32 years. Each subject was required to fill in a suitable anamnesis questionnaire in order to verify they did not undergo lower-limb injuries, disease, or trauma.

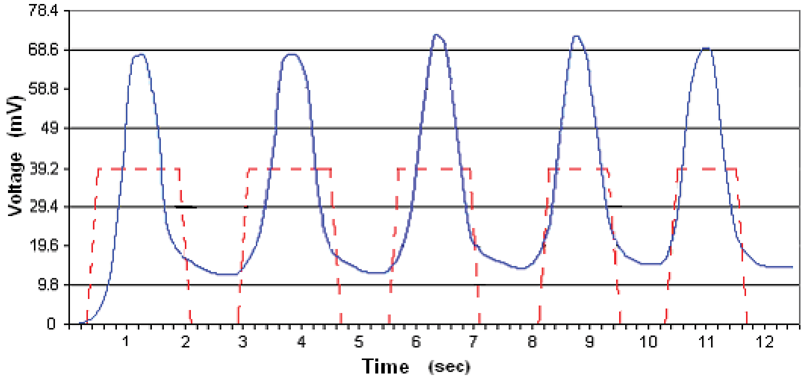
The "pathologic" sample consisted of five patients, four females and one male (mean age about 73 years), in-patient in a clinic specialized in vascular diseases, with special competences in cutaneous ulcers treatment (the medical partner in this work). The presence of the pathology under study was certified by consulting clinical folders.

After the measurement system was correctly applied, taking care of arranging the sensors upon the knee articulation, each subject was required to walk freely on a level ground. By means of a wireless communication system, based on Bluetooth protocol, data were transferred in real time from an on-body unit to a personal computer, where a software interface let the operator monitorize signals from the sensorized garment, including diagnostics and current settings. In a five-gait cycle observation interval the flexion-extension signals were acquired and the correspondent files stored for off-line elaboration.

DATA ANALYSIS

In the typical signal from a sensor of KB (continuous curve in Figure 4), resistance increases during flexion and decreases while the articulation extends, according to the piezoresistive effect. Voltage shows the same behaviour because the supply current was fixed for this kind of measurement. The synchronized matching with step-signal acquired from the sensorized shoe (dashed line) allowed us to detect the "gait cycle," or the elemental reference interval of the analysis. A single step-signal is an "on-off" signal where the low-level indicates that the foot is in contact with the ground, while the high-level indicates lifting.

Figure 4. Flexion-extension of the central sensor of KB (continuous curve) and signal coming from the shoe sensor (dashed curve) which marks the gait cycle



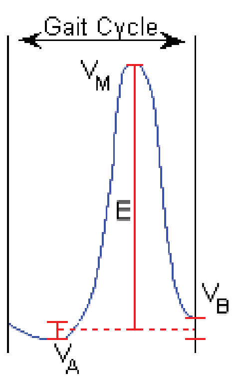
In order to extract significant variables for discrimination, two different approaches were studied. The first one, based on the evaluation of flexion-extension capability, was referred to as the *width excursion* parameter E, defined in Equation (1).

$$E = \left| V_M - \left(\frac{(|V_A - V_B|)}{2} + V_A \right) \right|. \tag{1}$$

V_M represents the maximum voltage width in the maximum flexion instant; V_A is the voltage value when flexion starts; V_B indicates the voltage value in the maximum extension instant such as shown in Figure 5. The second approach proposes to analyze gait discontinuity during the observed interval.

To do this we referred to the standard deviation over a sample made up of average width excursion (E) values, on the five-gait cycles. This parameter called *irregularity* (IR) was calculated both for central (IR_c) and lateral (IRL) sensors. In order to integrate information coming from each couple of sensors, a standard deviation was calculated with respect to their average value:

Figure 5. Graphical representation of the width excursion



$$E_{m_C} = \frac{E_{C_1} + E_{C_2}}{2}. \tag{2}$$

$$E_{m_L} = \frac{E_{L_1} + E_{L_2}}{2}. \tag{3}$$

E_{Ci} (i = 1, 2) indicates the width excursion regarding to central sensor i, while E_{Li} represents the width excursion related to lateral sensor i. IR parameters were computed by “nondistortion” method such as shown in Equation (4):

$$IR = \sqrt{\frac{\sum (x - \bar{x})^2}{(n-1)}}. \tag{4}$$

where x represents the sample arithmetical mean, while n is the dimension of the sample.

RESULTS

According to graphic representation of both IR_c and IR_L parameters over the entire recruited population (see Figure 6), several important issues have to be addressed:

- IR values showed fundamentally to be higher for pathologic subjects than for normal ones, confirming how they reveal an irregular knee flexion-extension during normal walking.
- A good discrimination between the two populations was obtained.
- IR values is higher in patients 5 and 1, which have a larger ulcer size, as it is known in clinical folder.
- Figure 6 reports on the discrimination on the basis of the irregularity of the mean value of central sensors IR_c and Figure 7 shows the graphical discrimination of subjects according to the irregularity of the mean value of lateral sensors IR_L. The graph IR_c reports a normal subject (normal 4) classified as belonging to the pathologic group, while in IR_L graph he is rightly recognized as normal. This means that, for our application, lateral sensors of KB resulted more specific than the central ones.

TEST EVALUATION

The test validity has to be intended as the capability by the test to discriminate between normal and pathological subjects within the recruited population. Receiver Operative Characteristic (R.O.C.) curves analysis provided its scientific support in this evaluation, mainly to decide the best cut-off value of parameters for discrimination. A R.O.C. curve is a graphic representation of sensitivity vs. false positive rate (FPR) for different cut-off levels of the measurement variable.

The best decisional value ensures maximum of sensitivity to the minimum FPR. By this method, optimal cut-off value of irregularity was detected both for central sensors and for lateral ones (see Figure 8) and the test characteristic parameters were computed, in terms of sensibility, specificity and positive predictive value (PPV), as shown in the Table 1. Positive predictive value is directly proportional to disease prevalence in the study population. For a good test a high PPV is more significant if prevalence is high too. In our study, prevalence was high enough (55%), so PPV values of 80% and 100% gave to the test a good validity score, together with high sensibility and specificity. Lateral sensors showed to be equally sensitive but better specific and predictive than central ones. However these results should be verified over a wider population.

Figure 6. Graphical representation of IR_c parameter

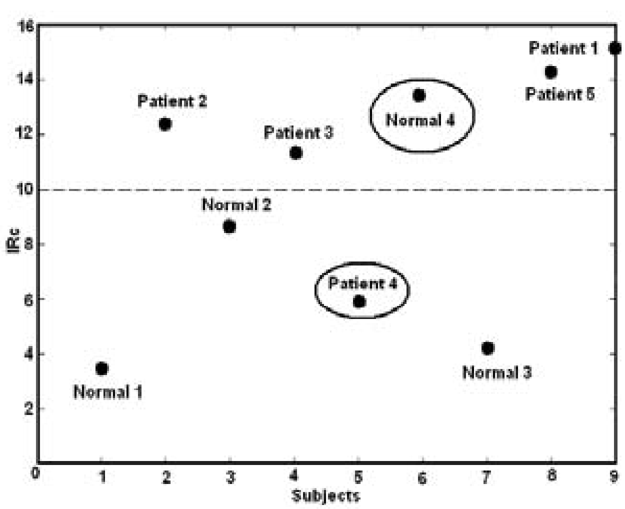
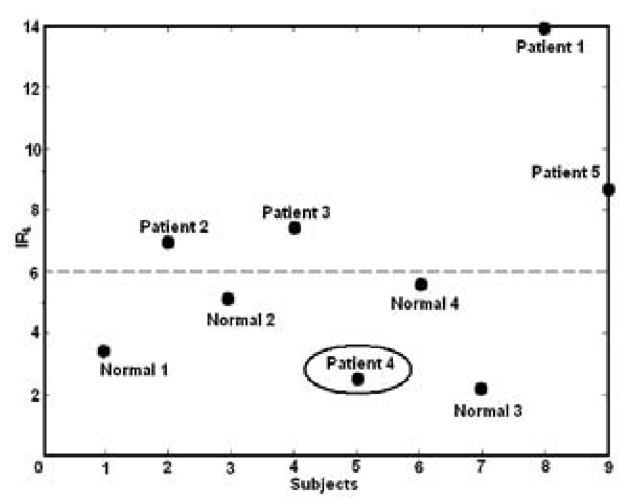


Figure 7. Graphical representation of IR_L parameter



CONCLUSION

In this article, we reported on the possibility of using a wearable kinesthetic system for monitoring through gait analysis the improvement of patients suffering from venous ulcers after undergoing surgical operation. As a validation tool, a preliminary pilot test has been realized, which reports a good capability of discriminating health from injured subjects. Furthermore, results demonstrated the possibility of identifying the severity of the disease in the pathologic group by analyzing of walking activity.

Commonly in the clinical practice, these patients are monitored by means of obtrusive instrumentation aimed at evaluating how the circumference of a thigh increases over time. We used an alternative approach to investigate the illness progress. As patients suffering from this illness exhibit an irregular gait, we customized a sensing garment in order to analyze the gait discontinuity.

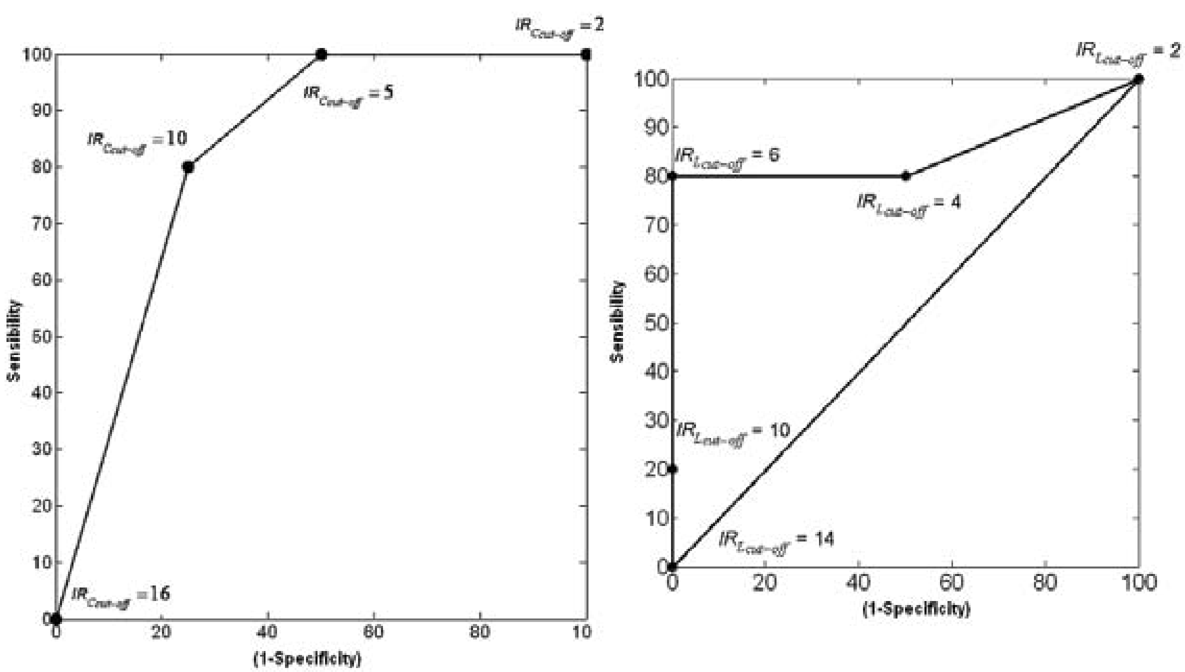
Therefore this prototype could be a means that allows clinicians to monitor patients without causing any discomfort during the recovery process after a surgical operation.

Next, developments aim at extending this application to the study of hip motion. Finally, it has been pointed out that the use of these sensorized garments can be considered a valid alternative and comfortable instrumentation applicable in several rehabilitation areas, such as in sport disciplines and in multimedia fields.

Table 1. Sensitivity, specificity, ppv and vpn value of the sensors central and lateral

Central sensors		**Lateral Sensors**	
Sensitivity	80%	Sensitivity	80%
Specificity	75%	Specificity	100%
PPV	80%	PPV	100%
VPN	75%	VPN	80%

Figure 8. Central sensor and Lateral R.O.C. curves



REFERENCES

Capozzo, A., Della Croce, U., Leardini, A., & Chiari, L. (2005). Human movement analysis using stereophotogrammetry part 1: Theoretical background. *Gait and Posture, 21*(21), 186-196.

De Rossi, D., Santa, A.D., & Mazzoldi, A. (1999). Dressware: Wearable hardware. *Ma. Sci. Eng. C, 7,* 31.

De Rossi, D., Lorussi, F., Scilingo, E. P., Carpi, F., Tognetti, A., & Tesconi, M. (2004). Artificial kinesthetic systems for telerehabilitation. In A. Lymberis & D. de Rossi, (Eds.), *Wearable e-health systems for personalised health management - state of the art and future challenges* (pp. 106-114). Amsterdam: IOS Press.

Lorussi, F., Rocchia, W., Scilingo, E.P., Tognetti, A., & De Rossi, D. (2004, December). Wearable redundant fabric-based sensors arrays for reconstruction of body segment posture. *IEEE Sensors Journal, 4*(6), 807-818.

Morris, S.J., & Paradiso, J.A. (2002). Shoe-integrated sensor system for wireless gait analysis and real-time feedback. In *Proceedings of the 2nd Joint Meeting of IEEE EMBS and BMES*, Houston, TX (pp. 2468-2469).

Post, E.R., Orth, M., Russo, P.R., & Gershenfeld, N. (2000). Design and fabrication of textile-based computing. *IBM System Journal, 39*(3-4).

Scilingo, E.P., Lorussi, F., Mazzoldi, A., & De Rossi, D. (2003). Sensing fabrics for wearable kinaesthetic-like systems. *IEEE Sensors Journal, 3*(4), 460-467.

Tognetti, A., Lorussi, F., Bartalesi, R., Tesconi, M., Quaglini, S., Zupone, G. et al. (2005, March). Wearable kinesthetic system for capturing and classifying upper limb gesture in post-stroke rehabilitation. *Journal of NeuroEngineering and Rehabilitation, 2*(8).

KEY TERMS

Anatomical Repere: A prominent structure or feature of the human body that can be located and described by visual inspection or palpation at the body's surface; used to define movements and postures. Also known as anatomical landmark.

Conductive Elastomer Composites: A rubberlike silicone material in which suspended metal particles conduct electricity.

Piezoresistive Effect: The changing electrical resistance of a material due to applied mechanical stress.

Quantitative Gait Analysis: Useful in objective documentation of walking ability as well as identifying the underlying causes for walking abnormalities in patients with cerebral palsy, stroke, head injury, and other neuromuscular problems

Wearable Systems: Devices that are worn on the body. They have been applied to areas such as behavioral modeling, health monitoring systems, information technologies, and media development.

Web-Enabled System Design for Managing Clinical Information

Leonor Teixeira
University of Aveiro, Portugal

Carlos Ferreira
University of Aveiro and Operational Research Center (CIO), University of Lisbon, Portugal

Beatriz Sousa Santos
University of Aveiro and IEETA, Portugal

INTRODUCTION

Healthcare is characterized by a highly complex environment where the process of patient care requires an unusual amount of communication between different healthcare professionals (HCPs). For a better patient care, the various HCPs have to cooperate, a processed often called shared care (Garde & Knaup, 2006; Schabetsberger, Ammenwerth, Andreatta, Gratl, Haux, & Lechleitner, 2006). Nowadays, there is an increasing incorporation of a heterogeneous set of Information Systems (ISs)—paper-based and computer-based—on the daily work of HCPs, in order to retrieve information about patients (Coiera, 2003; Van-Bemmel & Musen, 1997). The complexity of the patient care process combined with the heterogeneity of the information resources leads to a paradigm of data redundancy in the healthcare services in general, and hemophilia care in particular.

Hemophilia is a chronic disease that affects about 400,000 people worldwide; however, most of these people do not have access to adequate treatment (Evatt, 2005). A system for patient registry is a critical tool for monitoring, identification, and diagnosis of these patients; furthermore, it serves as an essential tool for managing their treatment. A registry is a database or a collection of records of people identified as having hemophilia or inherited bleeding disorders (Evatt, 2005). The purpose of a registry is to define the population demographics and collect observational data on specific hemophilia health concerns such as the prevalence of viral infections, factor inhibitors, implementation of prophylaxis for children or different product usage (Baker et al., 2004).

Portugal, in spite of having about 1,000 patients with hemophilia, doesn't have a hemophilia national patient registry, and most hemophilia treatment centers (HTCs) don't have a specific system to store and manage information concerning this pathology.

In order to help the management of this information at the *Hematology Service of Coimbra Hospital Center* (HS_CHC), as well as to facilitate communication between HCPs and patients, and improve the utility and quality of clinical data and treatment information, a Web-based IS under study is briefly presented in this work. To guide this study, namely along its design process, we followed an approach that combines a *grounded theory* approach with evolutionary design based on constant development, evaluation, and refinement of the generic domain model.

BACKGROUND: HEMOPHILIA CARE, INFORMATION AND TECHNOLOGIES

Hemophilia is an inherited bleeding disorder caused by low concentrations of specific coagulation factors (Bolton-Maggs & John-Pasi, 2003). With proper treatment, people with this disease can live healthy lives; otherwise, hemophilia causes crippling pain, severe joint damage, disability, and death (Baker et al., 2004).

Around 30 years ago, the development of coagulation-factor concentrates changed the life of these patients and allowed to treat them at home maintaining a high level of independence; thus, currently, most people with severe hemophilia are on therapy at home with infusion of the coagulation-factor concentrate. While some studies demonstrated significant quality of life benefits associated with home therapy using coagulation-factor concentrates (Teitel et al., 2004), this increased the difficulty to monitor their treatment and progress as well as to detect complications. For instance, many patients arrive at their care center without their home treatment record. Even when they have completed a paper record, uncertainties are often expressed as to the validity of the data, because it is possible that much of the data is entered retrospectively immediately prior to the clinic visit. Thus, there is an urgent need for the collection of meaningful outcome data for hemophilia, and in turn this means being able to obtain accurate records of home administration of coagulation-factor concentrates (Baker et al., 2004).

Web-based IS have the potential to support the ongoing care needs of patients as well as the collaborative management of with a chronic disease, like hemophilia. According to Baker et al. (2004) and Walker et al. (2004), one of the most relevant applications of electronic recording through the Web is the identification of early episodes of bleeding in the home setting, which otherwise would not have become apparent until the time of the patient's next clinic visit. A Web-based IS is also particularly helpful in identifying inappropriate use of coagulation-factor concentrates during home therapy.

Since hemophilia care is very expensive, cost-effective use of resources is extremely important; however, ensuring better data quality requires a continuous improvement process that includes the adoption of new information technologies, timely data entry, and rigorous audit and definition of data fields (Baker et al., 2004; Haux, 2006). Improving communication between patients and hemophilia center HCPs and the quality and timeliness of data collection and manipulation gives a better opportunity to improve long-term clinical outcomes in a cost-effective way (Collins et al., 2003). Moreover, improved communication pathways with HTCs can help minimizing the sense of isolation that patients with this disease may experience.

WEB-BASED INFORMATION SYSTEM TO SUPPORT HEMOPHILIA CARE

Motivation and Problem Context

The HS_CHC (one HTC in Portugal) requested the present study in order to evaluate the feasibility and usefulness of a Web-based IS to collect, record, and manage hemophilia patient data, based on the previous mentioned reasons. This HTC provides assistance to patients in three integrated Hospitals (*Central Hospital*, *Bissaya Barreto Maternity Hospital* and *Pediatric Hospital*) and provides clinical and laboratorial support to other hospitals all over Portugal.

Nowadays, the Portuguese Health Service uses several different computer-based ISs, in order to support the information flow and communication between different HCPs. The computer-based ISs used in HS_CHC are, basically: (i) *Integrated Hospital Information Systems* (IHIS)—which allow to visualize, manage and archive the administrative information while creating a clinical process; (ii) *Medical Support Information System* (MIS)—which allows to visualize, manage, and archive the clinical information during the medical appointment; (iii) *Nursing Support Information System* (NIS)—which allows to visualize, manage and archive nursing information during treatments; (iv) *ClinidataXXI*—to archive clinical analysis results and laboratorial information providing online service at the three aforementioned hospitals.

These computer-based ISs represent generic solutions, since they were developed to support general requirements of hospitals and cannot respond to the specific needs of hemophilia care. However, hemophilia center HCPs generate a lot of information when they observe their patients. Part of this information is in electronic format and is stored in computer-based ISs, but another part (i.e., home treatment records) is on paper format and is stored in paper files.

At present, patients record the result of hemophilia home therapy in paper diaries and send them to the HTCs by post, or bring them when they attend routine review appointments. This system has a number of weaknesses, as often paper records are incomplete or not returned. Furthermore, the period between individual infusions of treatments and receipt data may be long, which is undesirable, since that data is very important for clinical decisions about treatments. In this pathology, the information result of home therapy is more than a resource, since clinical decisions about patients' treatment depend on it.

The fact that the information concerning hemophilia patients is spread throughout different ISs is an obstacle for quality of data, fast retrieval of information, and thus healthcare quality. The need of an integrated IS for this specific pathology is obvious, and new technology has an important role in this field for developing a Web-based IS.

Our Proposal and Approach Followed in to Design

As mentioned, the goal of this proposal is to develop a Web-based IS to support the management of inherited bleeding disorders, which integrates, distributes, and archives large sets of information from heterogeneous sources in the scope of hemophilia care at HS_CHC.

To conduct this study, we followed an approach that combines a *grounded theory* approach with evolutionary design based on the constant development and refinement of the generic domain model, as shown in Figure 1.

Without constant and rigorous evaluation, the impact of new systems on the quality of care is unknown, and it is possible that badly designed systems could significantly harm patients (Cysneiros, 2002). To overcome these limitations we followed a specific approach to design the hemophilia Web-based IS. The main idea of our approach was the use of *grounded theory* to support the process of requirement engineering in a complex environment on the basis of a comprehensive domain model.

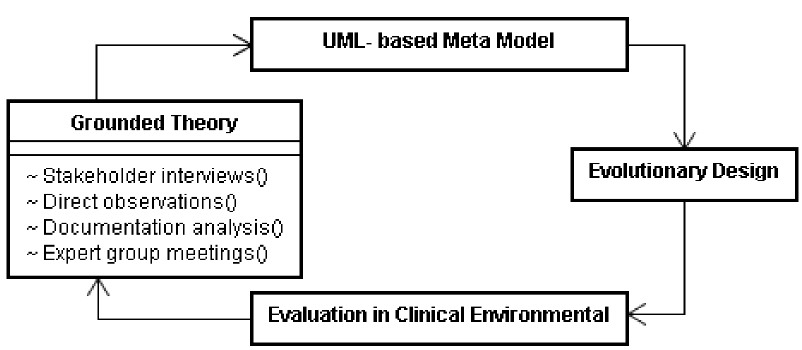
The *grounded theory* approach was published by Glaser and Strauss (1967) as a qualitative method in social research. It supports the inductive development of theories which are grounded in data to explain a phenomenon (Denzin & Lincoln, 2000; Strauss & Corbin, 1998), in the present context, the clinical practice in the HS_CHC. A theory is developed by systematically recording and analyzing data related to the phenomenon, in a process called constant comparison.

For the development of this domain model, we used the methods of software engineering, basically in two stages: requirement analysis and modeling. Requirement analysis is an important part of the software engineering process, where the needs or requirements of the system are identified (Maciaszek, 2005; Teixeira, Ferreira, & Sousa Santos, 2005). During the first step of the requirement analysis phase, we analyzed the clinical practice in typical environments of HS_CHC, in order to stimulate basic research, as well as understand the information flow and the management process of clinical data (Teixeira, Ferreira, & Sousa Santos, 2007). This analysis indicated that even though departmental information systems have sometimes been interfaced in order to transfer some data among a patient registration system, laboratory system, and medical system, this was usually based on ad-hoc activities without the implementation of a global communication concept.

In the second step of the requirement analysis, we tried to have a clear and detailed specification of the actors' functional requirements. The aim of this stage was to identify the characteristics of information, the users' needs, the key tasks they will perform with the system and the system functionality.

Basically, we started with a systematic analysis of the clinical practice, using direct observation, documentation analysis and unstructured interviews, in order to understand the particular characteristics of

Figure 1. High-level overview of the steps followed in design [adapted - Garde & Knaup, 2006]



this specific practice. Based on this, we developed a first domain model. In order to validate and improve it and to develop a valid theory (triangulation), this method was complemented with other gathering data techniques. We systematically collected, structured, and discussed the results in regular meetings with the expert physicians. From the experience and feedback of human actors we refined our domain model.

Having identified these requirements, appropriated modeling techniques had to be used. An object-oriented environment of design was selected, since it would allow to incorporate new components based, for instance, on decision support methods or techniques for formulating complex queries on the stored data. Thus, we applied the *Unified Modeling Language* (UML) to create a comprehensive domain model for capturing the requirements. We have chosen UML since it is an object-oriented design technique representing a standard language for specifying, visualizing, constructing, and documenting the artifacts of systems (Booch, Rumbaugh, & Jacobson, 1999; Whitten, Bentley, & Dittman, 2004).

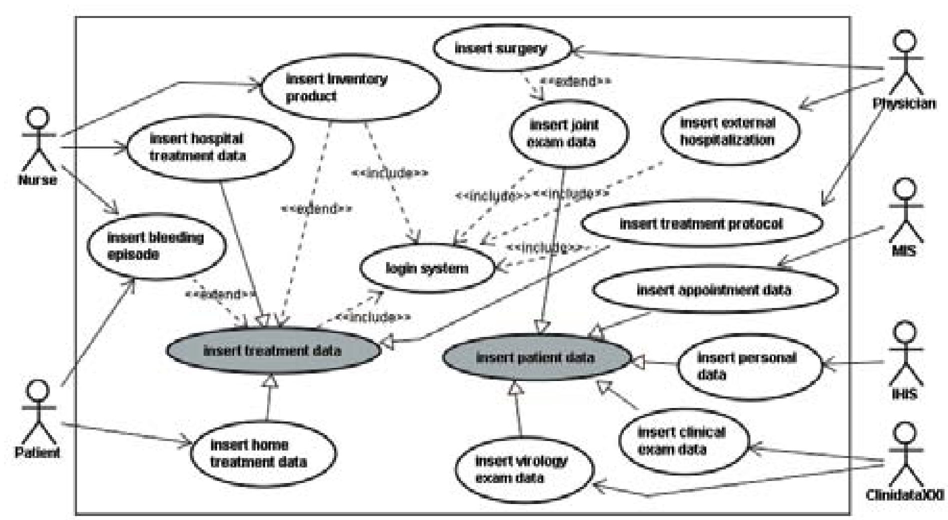
As a result of the requirement analysis step we identified the actors, functionalities of the system, and main classes of information for modeling the domain. We used the *use case* diagram of UML to describe external elements (actors) and main functionality (*use cases*); they are generally regarded as an excellent technique for capturing the functional requirements of a system (Fowler & Scott, 1997). The identified actors are the physician, nurse, patient (human users), and IHIS, MIS, ClinidataXXI (other systems). Figure 2 shows which actors interact with each *use case* for the main functionality, as well as the key processes involved in information introduction.

Patients have an individualized password protected home page, which can only be accessed by themselves or, in case of children, old people or illiterates, by a caregiver. For example, when a patient receive a treatment (infusion of coagulation-factor concentrate), they can access their personal page and enter details about treatment, symptoms of bleeding episodes associated, and the amount and batch number of coagulation-factor concentrate. If the treatment takes place in HCT, these data could be inserted by a nurse that has an individualized password too. The patient's page shows a recommended treatment regimen for prophylaxis and management of bleeding episodes. These treatment protocols are entered by a physician and can be updated at any time. Patient clinical evolution can be assessed trough the system in graphical and table formats. The HIS, MIS, and ClinidataXXI represent no-human actors and insert some data as personal data, clinical and virology exam data, and appointment data.

In order to detail the dynamic aspects of collaboration for the identified functionality, we used de sequence diagrams of UML. *Sequence* diagrams are interaction diagrams that specify how functionalities are carried out (what messages are sent and when).

For example, the diagram shown in Figure 3, was used to describe how objects (actor and classes) collaborate to execute the "insert treatment" functionality.

Figure 2. High-level overview of the UML use case diagram (including insertion tasks only)



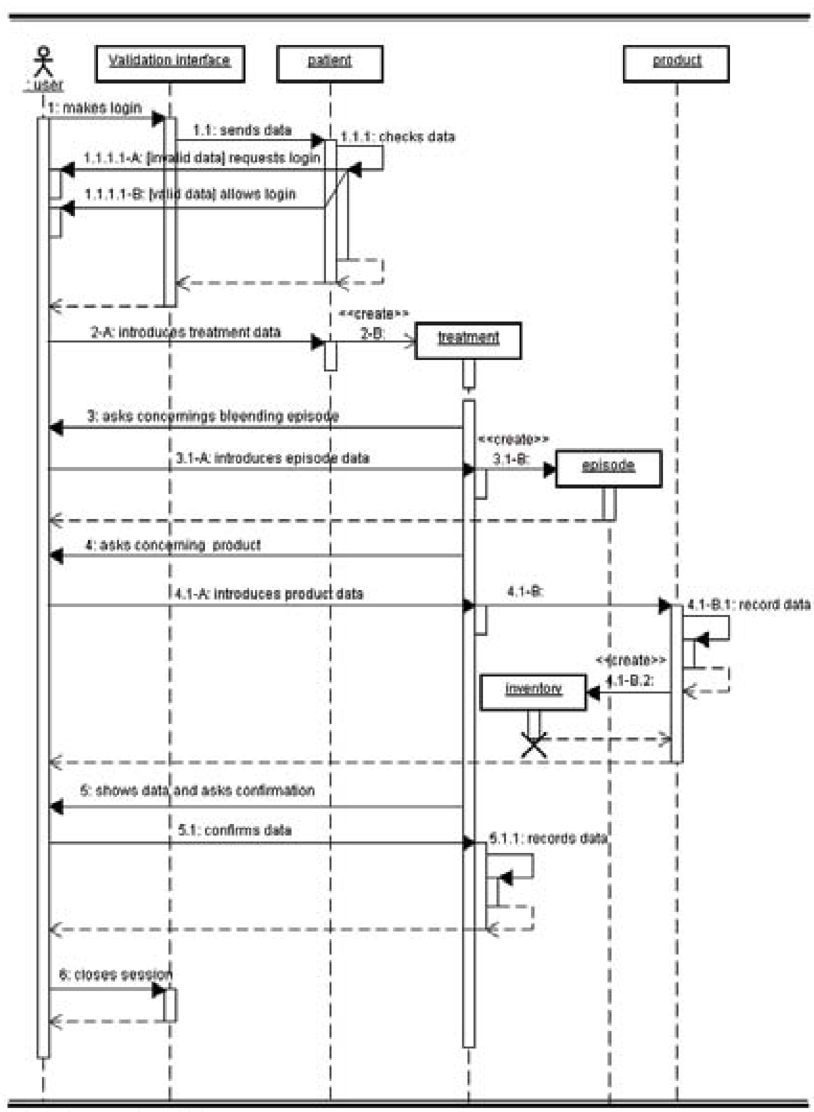
In order to represent the static structure showing object classes, attributes as well as relationships between those object classes, a *class* diagram of UML was used (Figure 4).

Besides patients' personal data, it is also important to record the data obtained from the medical program routines (appointments) and medical diagnosis. Patients' data are stored with a unique identification; each patient has a virology history, as well as a history concerning clinical analysis, external hospitalization, and so on. Also regarding the patient, it's necessary to store additional data obtained during the first medical appointment and routine review appointments, such as family tree and family history, joint condition, treatment history, and life quality.

Each patient can undergo many treatments. The treatments can be on-demand or prophylactic. On-demand treatments have associated bleeding episodes (in a joint, muscle, or mucosa). On the other hand, treatments can be made at home, at HTCs or at another hospital.

Usually, treatments consist of infusion of blood products (coagulation-factor concentrates), which are

Figure 3. UML sequence diagram of the "insert treatment" task



Note: user can be a patient or a nurse

Figure 4. UML class diagram (giving overview of domain model information)

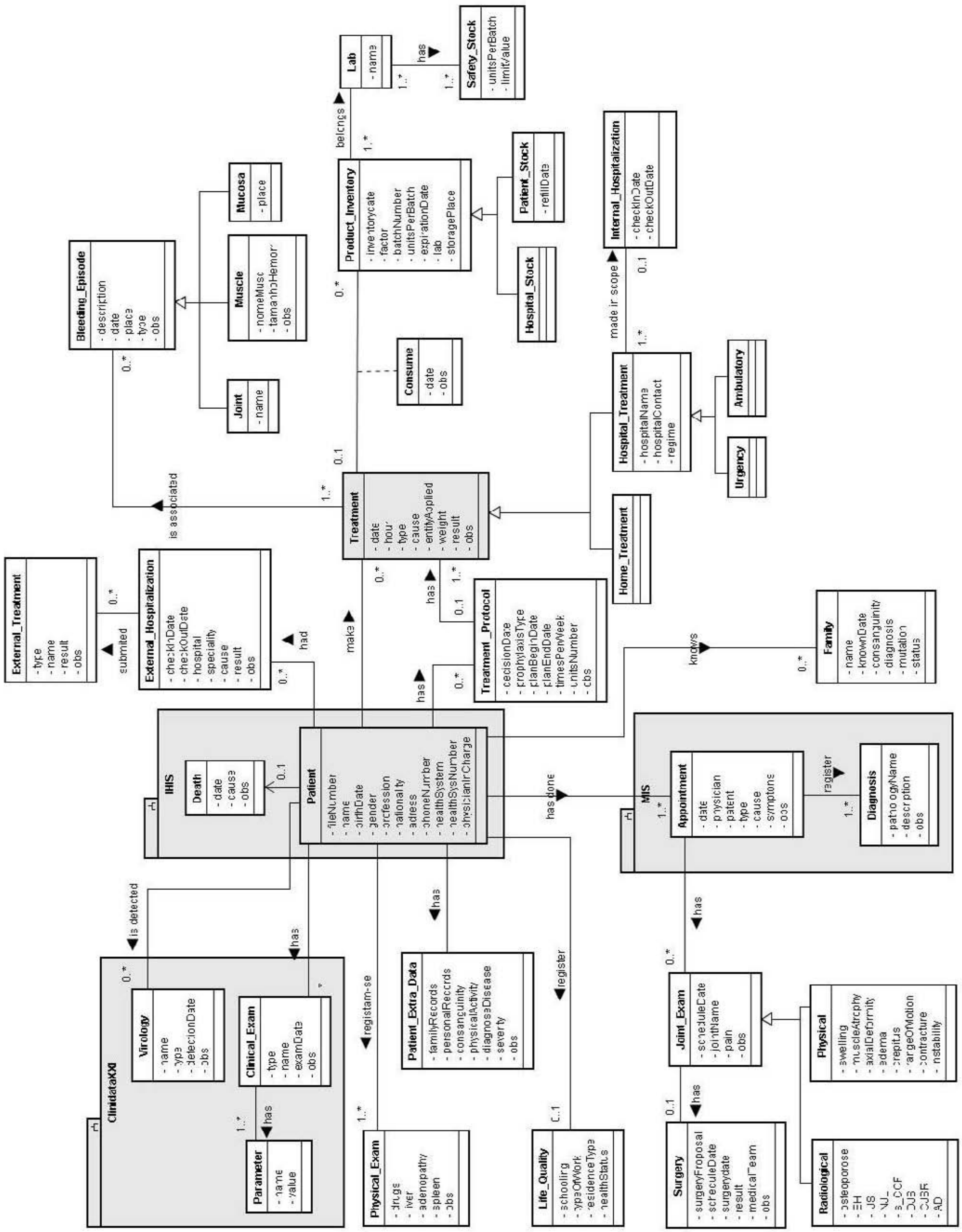


Table 1. Relationship between the functionality (presented in the use case diagram) information classes (presented in the class diagram) and corresponding permissions

Actor	Funcionality	Information Classes (permission)
Patient	+ insert treatment data - home treatment + insert bleedind episode	+ Patient (read) + Treatment (write) + Product_Inventory (update) + Bleeding_Episode (write)
Nurse	+ insert treatment data - hospital treatment + insert bleedind episode + insert product inventory	+ Patient (read) + Treatment (write) + Bleeding_Episode (write) + Product_Inventory (write, update)
Physician	+ insert patient data - joint exam + insert surgery + insert ext hospitalization + insert treatment data - insert treat protocol	+ Patient (read) + Family (write, update) + Diagnosis (write) + Physical_Exam (write) + Patient_Extra_Data(write) + Life_Quality (write, update) + Joint_Exam (write) + External_Hospitalization (write) + Treatment_Protocol (write, update)
MIS	+ insert patient data - appointment	+ Appointment (write) + Diagnosis (write)
IHIS	+ insert patient data - personal data	+ Patient (write)
ClinicadaXXI	+ insert patient data - clinical exam - virology exam	+ Clinical_Exam (write) + Virology (write)

Note: write permission receives by inheritance read permission.

in stock; thus, functionality concerning the management of the blood products inventory should also be integrated. When a blood product arrives at HTC, a nurse inserts its data into the system (name of product, data of reception, batch number, number of units per batch, validity); on the other hand, a blood product can be in different places, as it can be given to patients for home treatment or to a distant treatment center. When a blood product is consumed in the sequence of a treatment, the system should register this occurrence, and automatically update the stock.

All this information can be visualized (*read*), inserted (*write*), and updated (*update*) by different actors through different functionality (as shown in Table 1).

FUTURE TRENDS

We are currently implementing, for the *Hematology Service of Coimbra Hospital Center*, a prototype of a Web-based IS according to the above described domain model. If this Web-based IS proves to be a good tool for delivering hemophilia information services in this center, we would like o extend this solution to other HTCs in Portugal, in order to contribute to the development of a hemophilia national patient registry. Moreover, we should also take into consideration that there are significant technical and regulatory issues surrounding the privacy, security, and confidentiality of individually identifiable health information, which need to be settled.

CONCLUSION

The use of electronic records within healthcare systems is increasing and the quality, usefulness and availability of patient information is potentially improved by this approach. Particularly in the hemophilia care, the availability of specialized IS within this community is gaining more importance.

Due to the decentralization of the staff working places, an efficient and effective integrated system can provide the necessary link between data collection, storage, and analysis. The ISs together with the Internet offer presently significant opportunities to HCPs and patients to improve their communication and joint management of chronic diseases, as hemophilia, at the same time, they offer to HCPs analytical tools to transform data into information and mine relevant information through data queries. On the other hand, using a Web-based IS, patients can have direct access to the system, to view their clinical history, as well as insert treatment data. The management and stock control of the products used in treatments can also be improved.

In this work, we described the design process of a Web-based IS for managing clinical information in hemophilia care to support patients assisted at the *Hematology Service of Coimbra Hospital Center*.

In order to collect the data and gather knowledge about domain of the *Hematology Service of Coimbra Hospital Center*, a *grounded theory* approach was conducted using techniques such as interviews, direct observations, expert group meetings, and analysis of existing documentation. The first phase was to identify the information gaps in the service. Then, we proceeded to the requirement analysis with business process assessment, direct, and indirect observation techniques and interviews with HCPs. Six actors were identified: patient, physician, nurse, and three already-existing ISs (MIS, IHIS and ClinidataXXI). Based on the identified actors and needs, a generic conceptual model was developed using *Unified Modeling Language*.

Healthcare industry is characterized for highly complex environments and information technologies can have a significant influence on the patient's health. However, technology misconception and/or misuse can be risky, and in severe cases, it can lead to the patient death. For this reason, any new systems have to be carefully validated and verified before becoming a part of routine patient care. For this reason, a good design process can give a major contribution, and we believe that by using a *grounded theory* based approach for the requirements engineering phase, we were able to design a Web-based IS, we deem will fulfill the needs of the Hemophilia Service. A prototype is currently under development and has already passed a first evaluation cycle.

ACKNOWLEDGMENT

The authors are grateful to Doctor Natália Martins of *Hematology Service of Coimbra Hospital Center* for providing us access to data and information systems, as well as for all the help in the requirement analysis and discussion of the proposed model.

REFERENCES

Baker, R., Laurenson, L., Winter, M., & Pritchard, A. (2004). The impact of information technology on haemophilia care. *Haemophilia, 10*(4), 41–46.

Bolton-Maggs, P., & John-Pasi, K. (2003). Haemophilias A and B. *Lancet, 361*, 1801– 1809.

Booch, G., Rumbaugh, J., & Jacobson, I. (1999). *The unified modeling language guide*. Reading, MA: Addison-Wesley.

Coiera, E. (2003). *Guide health informatics* (2nd ed.). London: Hogder Arnold.

Collins, P., Bolton-Maggs, P., Stephenson, D., Jenkins, B., Loran, C., & Winter, M. (2003). Pilot study of an Internet-based electronic patient treatment record and communication systems for haemophilia, Advoy.com. *Haemophilia, 9*(3), 285–291.

Cysneiros, L. (2002). Requirements engineering in the healthcare domain. In *Proceedings of 10th Anniversary IEEE Joint International Conference on Requirements* (pp. 350–356). Engineering, WA.

Denzin, N. K. (2000). *Handbook of qualitative research* (2nd ed.). Thousand Oaks, CA: Sage.

Evatt, B. (2005). *Guide to developing a national patient registry*. Québec, Canada: World Federation of Hemophilia.

Fowler, M., & Scott, K. (1997). *UML distilled - applying the standard object modeling language*. Reading, MA: Addison-Wesley Longman, Inc.

Garde, S., & Knaup, P. (2006). Requirements engineering in healthcare: The example of chemotherapy planning in paediatric oncology. *Requirements Engineering, 11*(4), 265–278.

Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory: Strategies for qualitative research.* Chicago, IL: Aldine Publishing Company.

Haux, R. (2006). Health information systems - past, present, future. *International Journal of Medical Informatics, 75*(3-4), 268–281.

Maciaszek, L. (2005). *Requirements analysis and systems design* (2nd ed.). Addison Wesley.

Schabetsberger, T., Ammenwerth, E., Andreatta, S., Gratl, G., Haux, R., Lechleitner, G., et al. (2006). From a paper-based transmission of discharge summaries to electronic communication in healthcare regions. *International Journal of Medical Informatics, 75*(3–4), 209–215.

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (2nd ed.). Newbury Park, CA: Sage.

Teitel, J., Barnard, D., Israels, S., Lillicrap, D., Poon, M.-C., & Sek, J. (2004). Home management of haemophilia. *Haemophilia, 10*(2), 118–133.

Teixeira, L., Ferreira, C., & Sousa Santos, B. (2005). Uma Abordagem ao Processo de Desenvolvimento dos Sistemas de Informação: cuidados a ter ao longo do processo no caso dos SI's tradicionais e SI's distribuídos na Web. *Revista Iberoamericana de Sistemas, Cibernética e Informática, 1 (in Portuguese)*.

Teixeira, L., Ferreira, C., & Sousa Santos, B. (2007). Especificação e análise de uma Aplicação Web para a o Registo Electrónico do Paciente na área da Hemofilia – caso prático. In *Proceedings of Conferência Associação Portuguesa de Sistemas de Informação - CAPSI*, Aveiro – Portugal, 407–411, *(in Portuguese)*.

Van-Bemmel, J. H., & Musen, M. A. (1997). *Handbook of medical informatics*. Springer.

Walker, I., Sigouin, C., Sek, J., Almonte, T., Carruthers, J., Chan, A., et al. (2004). Comparing hand-held computers and paper diaries for haemophilia home therapy: A randomized trial. *Haemophilia 10*(6), 698–704.

Whitten, J. L., Bentley, L. D., & Dittman, K. C. (2004). *Systems analysis and design methods* (6th ed.). New York: McGraw-Hill/Irwin.

KEY TERMS

Grounded Theory: Research method most often associated with the social sciences.

Healthcare Actors: Individual or group with an interest in the healthcare system, including patients and their families, nurses, physicians, laboratorial technical staff, and other external entities as regulators, insurance companies, and healthcare organizations.

Home Therapy: Process where the hemophilic patients receive treatment with coagulation-factor concentrates in their homes.

Modeling Process: Definition of the details of a business process flow, modeling all the data, and other relevant elements.

Requirement Analysis: All tasks that go into determining the requirements of a new or altered system, taking in account users' needs.

Requirement Engineering: Systematic requirement analysis.

Shared Care: Continuous, patient-oriented cooperation of hospitals, general practitioners, specialists, and other healthcare professionals (HCPs) during patient care.

Software Engineering: Application of a systematic, disciplined, quantifiable approach to the development of Information Systems/Software.

Unified Modeling Language (UML): A standardized specification language for object modeling that includes a graphical notation used for visualizing, specifying, constructing, and documenting the artifacts of a system.

Web-Based Information System: Information System that is accessed with a Web browser over a network such as the Internet or an intranet.

Wireless Sensor Network to Support Home Care in Telemedicine Applications

Silvia Jiménez-Fernández
Grupo de Bioingeniería y Telemedicina,
Universidad Politécnica de Madrid, Spain

Antonio Cobo-Sánchez-de-Rojas
Grupo de Bioingeniería y Telemedicina,
Universidad Politécnica de Madrid, Spain

Álvaro Araujo-Pinto
Laboratorio de Sistemas Integrados,
Universidad Politécnica de Madrid, Spain

Pedro Malagón
Laboratorio de Sistemas Integrados,
Universidad Politécnica de Madrid, Spain

Octavio Nieto-Taladriz
Laboratorio de Sistemas Integrados,
Universidad Politécnica de Madrid, Spain

Paula de Toledo
Grupo de Bioingeniería y Telemedicina,
Universidad Politécnica de Madrid, Spain

Francisco del Pozo
Grupo de Bioingeniería y Telemedicina,
Universidad Politécnica de Madrid, Spain

INTRODUCTION

The care of patients suffering from chronic diseases is a growing source of expense for health care services around the world. The implementation of new models for patients' treatment and follow-up needs to be faced in order to increase patients' quality of life, and to reduce the costs associated. In this article, we propose a point-of-care for home care scenarios that is based on the remote monitoring of biomedical parameters.

BACKGROUND

Research experiences in the past have proved that home care can maintain or even improve old or chronically ill people's quality of life, as well as decrease health care costs associated. Home care delivery can still be improved through advances in human-machine interfaces that enable users (i.e., mostly old people) to deal easily and nonintrusively with technology (Ades, 2000; Hernandez, Casas, Escarrabill et al., 2003; Morlion, Knoop, Paiva, & Estenne, 2002). Many limitations brought on by aging are expected to be overcome by innovation in the area of Ambient Intelligence (AmI) (Cucchiara, Grana, Prati, & Vezzani, 2005; Trivedi, Huang, & Mikic, 2005), and focused on the design of devices easy to use and easy to learn that work transparently and pervasively to aid the user (Remagnino & Foresti, 2005).

Therefore, next generation home care and follow-up systems have important challenges to face, such as the development of nonintrusive systems that provide the right medical services on a highly-usable basis or systems that require minimum human-machine interaction to promote an augmented environment while making computers disappear in the background. Of course, it is important to make these necessities possible by means of developing affordable solutions for public and private health care institutions, only achievable with low cost structures easy to install and to maintain.

In order to achieve these requirements, we propose a Wireless Sensor Network, based on a Home Area Network (HAN), featuring a gateway, several biomedical sensors (pulse oximeter, electrocardiograph, and accelerometer to quantify daily activity) that the doctor recommends to the patient to control his/her health and a user interface, key element to inform the user of events registered (such as start of the exam, low battery supply, etc.). All network nodes are provided with wireless capabilities to allow full mobility (ISTAG, 2002).

THE BIOMEDICAL WIRELESS SENSOR NETWORK

Requirements

The requirements taken into account when designing the home care point-of-care have been mainly: self-management, context awareness, plug-and-play features, and security. Due to the need to achieve wearable and low cost sensors, limited resources in processing power have also been taken into account.

Regarding self-management and plug-and-play capabilities, the point-of-care has been designed to automatically route data packets. That is to say that the user does not have to consider nodes' breakup, or the addition of new nodes/sensors to the network, as they are transparently recognized and managed. For security reasons, data coming from new sensors is not sent to the telemedicine server until approved by it.

The nodes of the wireless network have to be context aware to adapt their behavior accordingly to environmental variables that govern the system, such as the actor of the system and the actor's preferences, the moment of the day when usage takes place, season, and so forth.

Once again, it is important to achieve these goals by means of low cost systems, as these points-of-care will not be deployed massively or used in clinical routine if they are not cost-efficient and provide benefits over traditional health care delivery. Of course, as costs are reduced, limitations arise, such as processing power, tight memory sizes, or low power consumption (batteries are required to last long enough to allow several measures to take place, as well as to support the wireless radio link power consumption).

Above all, there are security issues to be considered. For real scenario setups where sensitive and personal data will be in transit through the network, refusing intrusions, avoiding denial of service, driving the system into a safe state in case of failure, and so forth, are a must.

Functionality

Based on the requirements stated above, the wireless sensor network consists of three different elements: a gateway (center of the point-of-care), the sensors (used to monitor the patient's vital signs), and the user interface (to inform of upcoming exams, problems, etc.).

The functionality of the different elements follows. The gateway has twofold capability: (1) it is used to communicate with the sensors that the patient has, and (2) it is needed to send data to and from the health care institutions (i.e., hospital, primary care, etc.). The information regarding the doctor's prescription of medical exams is sent to the gateway where it is stored, processed, and routed to the specific sensor, stating the duration, sample rate, quality of the measure, amplification factor, or any precise parameter needed to collect the data.

The sensors have been developed as stand alone wireless devices and have been provided of automatic start of the measure, auto-discovery, and auto-configuration facilities. That is to say that the patient does not need to take care of configuring nor connecting them, for instance, to upload or download the data collected. It is important to note that the user does not need to ensure either that the sensors are powered on and off, as they are suspended and resume automatically after a given time. When resume is done every half minute, several tasks take place:

1. Checking the state of the sensor to become aware of the patient's willingness to start an exam. When the user wants to commence it, the user places the sensor on, so that the device is able to collect data. This situation takes place no matter whether the doctor programmed the test or the patient decided to take it (because of not feeling alright).
2. Connecting to the gateway. The radiofrequency module is powered on and a "keep alive" message is sent to the gateway. If there is information that needs to be transmitted to the node, as imminent exams that need to take place, it is sent at this moment. Afterwards, if data are still stored in the sensor, the data are forwarded to the gateway.
3. Checking battery level. When batteries are close to being used up, an alarm is sent to the gateway and the user is notified so that the user can recharge them. This check is performed every 10 minutes instead of every half minute.
4. Checking network failures.

Once all these tasks have been performed, the device is suspended again to save batteries. Using this method, battery consumption is decreased up to a maximum of 60%.

Results

The abovementioned approach has led to the definition and development of a home point-of-care that aims to foster independent living of the elderly or chronically ill featuring high usability and minimum human machine interaction and remote monitoring.

This has been achieved using a gateway to route information from the house to outside institutions (hospitals, private institutions, etc.), and also within the home environment, through a network of monitoring sensors. In order to minimize maintenance and installation costs, as well as to improve usability, a wireless network (LR-WPAN) has been used to interconnect the sensors and the hub (Jimenez-Fernandez et al., 2006).

Each node (i.e., sensing device) has three different parts. First, the biomedical sensor itself: an off-the-shelf OEM board (ECG, pulse oximeter) or an ad-hoc module (accelerometer) that has been provided with a second component, an autonomous power supply. As the power requirements of each sensor are different, customized supply stages have been designed using a standard interface with commercial AAA batteries, and always looking for the highest efficiency that maximizes sensors' life time. Finally, the third part that has been added provides the "intelligence" explained in the requirements section: awareness of patient's willingness to perform a measurement, network functionality (e.g., routing, reconfiguration), and so forth. The main component of this stage is a radio transceiver in the ISM band of 2.4 GHz that implements the physical layer and the MAC layer of the IEEE 802.15.4 wireless standard. The main characteristics featured are:

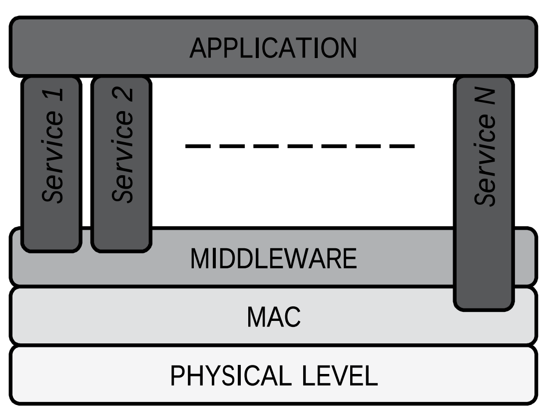
- Possible bit rates: 250 Kbps, 40 Kbps or 20 Kbps. We are using 250Kbps.
- Possible network topologies: mesh, star, or point to point. We are using star topology because it is more efficient for this propose.
- Support for low latency devices.
- Channel access using CSMA-CD (Carrier Sense Multiple Access/Collision Detection).
- Low power consumption.

Therefore, our network requirements and functionality suit perfectly with the given standard. In order to implement the remaining layers of the protocol stack, a low-cost PIC 16F microcontroller that connects to the radio transceiver through a SPI communication interface has been attached to each node. The commonly used OSI architecture has been refused in favor of an architecture based on services. The goal pursued with this design is the implementation of applications that arise as a mere aggregation of services and relations between services, rather than a node-focused orientation, simplifying the development of user applications (Sgroi, Wolisz, Sangiovanni-Vincentelli, & Rabaey, 2003). Our services are built on top of a communication middleware (Coleri, Puri, & Varaiya, 2003; Heinzelman, Murphy, Carvalho, & Perillo, 2004; Ranganathan & Campbell, 2003) that abstracts network or hardware specific tasks to the application. This middleware allows sending data to any node of the network without considering its location, sending broadcast messages, and so forth, and all this happens using a simple and unique interface (based on query-response) that can be dynamically instantiated in any node at any time. This allows us to easily compose applications taking into account many complex aspects, such as privacy, safety, load balancing, low energy consumption, and so forth (Coleri et al., 2003; Rabaey et al., 2000), and in a totally transparent way to the user. The protocol stack can be depicted in Figure 1.

The idea behind this approach is to isolate common behavior that can be reused from other applications and to encapsulate it as system services, so that once the system is running the addition of a new application becomes easy.

Another important fact that has to be emphasized is that, in contrast to traditional architectures, service oriented architectures do not force all services to access all middleware/layers available (as can be seen

Figure 1. Protocol stack



in Figure 1). For instance, in order to check battery levels, communication protocols can be bypassed and access to the physical level directly used, making signal processing more efficient.

All in all, an interoperable, reliable system, abstracting the hardware from the applications, has been designed, developed, and implemented. It is important to highlight that the size of the prototype is not comparable to that of wearable devices, but the aim of this work was just to emulate the behavior that wearable sensors will have in the future, not to miniaturize existing ones.

FUTURE TRENDS

This article has presented a system that addresses usability issues, plug-and-play interoperability, reconfigurability, and wireless communication between the different devices of the home's point-of-care. Emerging trends on the field of research are focused on applying standards to these issues in order to facilitate vendor interoperability and so avoid closed solutions. At the moment, there is a family of standards, ISO11073/IEEE 1073, that aims to increase interoperability at the hospital's point-of-care. Although it is not specifically designed to undertake the problem of integrating monitoring devices in the patient's house, it may be useful in our environment.

All in all, the aforementioned approach aims to solve the difficulties of device replacements and upgrades that usually lead to reconsidering the design of the whole system or, at least, to the modification of important software coding or hardware development. By applying these standards, communications between the different devices and the gateway will be transparent and universal no matter whom the vendor is.

CONCLUSION

The point-of-care described hereby has focused on usability concepts, aiming to develop intelligent and intuitive nodes able to respond to the activities and situations recognized in the user's environment.

The primary objective of this platform is to examine whether a system of basic biomedical sensors (motion sensors, ecg, pulse oximetry, blood pressure, etc.) is able to detect behavioral patterns and help aging and ill populations with everyday activities and health care follow up; that is to say, to provide the foundation for an Ambient Intelligence approach.

Finally, a flexible, dynamic, and easy to program architecture has been implemented, enabling software developers to focus on application level without paying attention to network and hardware levels.

REFERENCES

Ades, P.A., Pashkow, F.J., Fletcher, G., Pina, I.L., Zohman, L.R., & Nestor, J.R. (2000). A controlled trial of cardiac rehabilitation in the home setting using electrocardiographic and voice transtelephonic monitoring. *American Heart Journal, 139*(Suppl. 3), 543-548.

Coleri, S., Puri, A., & Varaiya, P. (2003). Power efficient system for sensor networks. In *8th IEEE International Symposium on Computers and Communication Proceedings (ISCC 2003).*

Cucchiara, R., Grana, C., Prati, A., & Vezzani, R. (2005). Probabilistic posture classification for human-behavior analysis. *Systems, Man and Cybernetics, Part A, IEEE Transactions, 35*(Suppl. 1), 42-54. ISSN: 1083-4427.

Heinzelman, W.B., Murphy, A.L., Carvalho, H.S., & Perillo, M.A. (2004). Middleware to support sensor network applications [Special issue]. *IEEE Network Magazine.*

Hernandez, C., Casas, A., Escarrabill, et al. (2003). Home hospitalization of exacerbated COPD patients. A randomized controlled trial of clinical efficacy and costs of services for chronically ill patients. *ERJ, 21*(Suppl. 1), 58-67.

Information Society Technologies Advisory Group. (2002). *STAG, Scenarios for Ambient Intelligence in 2010.* Retrieved February 6, 2008, from http://www.cordis.lu/istag.htm

Jimenez-Fernandez, S., Araujo-Pinto, Á., Cobo-Sánchez de Rojas, A., del Pozo-Guerrero, F., Nieto-Taladriz, O., de Toledo-Heras, P., et al. (2006). PERSEIA: A biomedical wireless sensor network to support healthcare delivery for the elderly and chronically ill. In *28th IEEE EMBS Annual International Conference* (pp. 2064-2066). ISBN: 14244-0033-3, ISSN: 1557-170X.

Morlion, B., Knoop, C., Paiva, M., & Estenne, M. (2002). Internet-based home monitoring of pulmonary function after lung transplantation. *American Journal of Respiratory and Critical Care Medicine, 165*(Suppl. 5), 694-697.

Rabaey, J.M. et al. (2000). PicoRadio supports ad hoc ultra-low power wireless networking. *IEEE Computer, 33*(7), 42-48.

Ranganathan, A., & Campbell, R.H. (2003). A middleware for context-aware agents in ubiquitous computing environments. In *Proceedings of the International Middleware Conference* (LNCS 2672, pp. 143-161). Heidelberg, Germany: Springer Berlin.

Remagnino, P., & Foresti, G.L. (2005). Ambient intelligence: A new multidisciplinary paradigm. *Systems, Man and Cybernetics, Part A, IEEE Transactions, 35*(Suppl. 1), 1-6. ISSN: 1083-4427.

Sgroi, M., Wolisz, A., Sangiovanni-Vincentelli, A., & Rabaey, J.M. (2004). *A service-based universal application interface for ad-hoc wireless sensor networks whitepaper.* Berkeley, CA: U.C. Berkeley.

Trivedi, M.M., Huang, K.S., & Mikic, I. (2005). Dynamic context capture and distributed video arrays for intelligent spaces. *Systems, Man and Cybernetics, Part A, IEEE Transactions, 35*(Suppl. 1), 145-163. ISSN: 1083-4427.

KEY TERMS

Embedded System: Special-purpose system in which the computer is completely encapsulated by or dedicated to the device or system it controls.

Home Area Network (HAN): The interconnection of information technology devices within the range of a house.

Home Care: Health care given to a person at home that aims to enable people to remain at home rather than use residential, long-term, or institutional-based nursing care.

LR-WPAN: Low Rate Wireless Personal Area Network.

Point-of-Care: Care at the bedside or anywhere else where the patient might be, or where the caregiver makes the decisions.

Usability: Perception of a target user of the effectiveness (fit for purpose), efficiency (work or time required to use), and ease of use of a device, interface, computer program, Web site, and so forth.

Wearable Sensors: Biomedical sensors that, due to their small size and weight, can be worn on the body.

Wireless Communication: Communication that takes place via airwaves as opposed to cables.

Workflow Management Systems for Healthcare Processes

C. Combi
Università degli Studi di Verona, Italy

G. Pozzi
Politecnico di Milano, Italy

INTRODUCTION

Organizations achieve their goals by executing processes, which can be physical processes to produce goods (e.g., assembling components of a PC), information processes to manage information (e.g., to collect information about facts), or business processes (e.g., to produce goods, to provide services, and to manage information). Final goals are achieved by executing atomic work units (tasks), by coordinating them, by assigning them to executors (agents) that can be human or automatic ones.

A workflow management system (WfMS) is a software system that supports the execution of such processes inside organizations. A WfMS requires a formal description, known as schema or process model or workflow, for every process it has to manage: many instances (or cases) of a process can be executed at the same time, sharing the same schema, while each case has its specific data, like patient's name, insurance claim, and so on.

A WfMS has to manage a huge amount of data: to do that, it relies on a Database Management System (DBMS), which is mostly based on the relational data model. A WfMS is also tightly coupled to the Information Systems (IS) of the organization, and may be even more tightly coupled to the Decision Support System (DSS) of the same organization. In fact, the IS mainly focuses on data and data flows, without considering in detail the processes that produce/manage/consume data; the DSS, instead, helps the management to identify strategies for the organization, considering aggregated data possibly stored in a data warehouse.

At the early 2000s, the term of "scientific workflow" has also been introduced, mostly in the fields of bioinformatics and chemical informatics. However, scientific workflows do not refer to the specific context of processes and activities that can be suitably managed by a traditional WfMS, as defined above: scientific workflows rather refer to the need for multiple interconnected tools to handle large quantities of data, like those coming from different monitoring systems, with multiple data formats, and for interoperability purposes.

BACKGROUND

A WfMS supporting the coordinated execution of tasks by agents requires (Grefen, Pernici, & Sanchez, 1999) a formal description of the process (process model), a formal description of the organization where the process is executed, and whose agents will be assigned for the execution of tasks (organizational model), and a formal description of the information related to every case (information model).

The workflow designer is responsible for the definition of the process model, generally performed via graphical design tools, producing the formal description in a Process Definition Language (PDL) format. Every WfMS features its own PDL. The Workflow Management Coalition – WfMC – (http://wfmc.org), which is a nonprofit organization involving producers, vendors, and consultants, as well as users of WfMSs, defined a standard process definition language, namely X-PDL, based on the XML format of documents: recently, some WfMSs and some Workflow Designer Tools have been released, declaring themselves to be X-PDL compliant.

The organizational model describes the structure of the organization where the process is executed, and considers agents available inside that organization. Typically, information stored for every agent, beyond traditional ones, relate to the skills of the agent (e.g., secretary, nurse, physician, head of the department, CEO), authorizations granted to the agent (e.g., responsibility for a surgery intervention), localization of the agent (e.g., NIH at Bethesda, MD), working hours, and so on: these information are relevant for the scheduling and assignment of tasks to agents, balancing the workloads among available agents and maintaining constraints of tasks (e.g., deadlines for the completion of a task or the need of having two distinct executors for two tasks, like when one agent makes a proposal and a different ones approves or rejects).

The information model describes the data that are relevant for the process itself, like the patient's name, the patient's choices, the preferred food or diet (e.g., vegetarian) and so on: every case owns its data, which thus are case-specific. Data from the information model are also used to identify execution paths inside the process model. For example, if the patient suffers from diabetes, the task PrepareSugarFreeMeal will be executed; otherwise, the task PrepareNormalMeal will be executed. Data from the information model are stored by the underlying DBMS and, in some cases, can be shared with the IS and with the DSS.

The Architecture of a WfMS

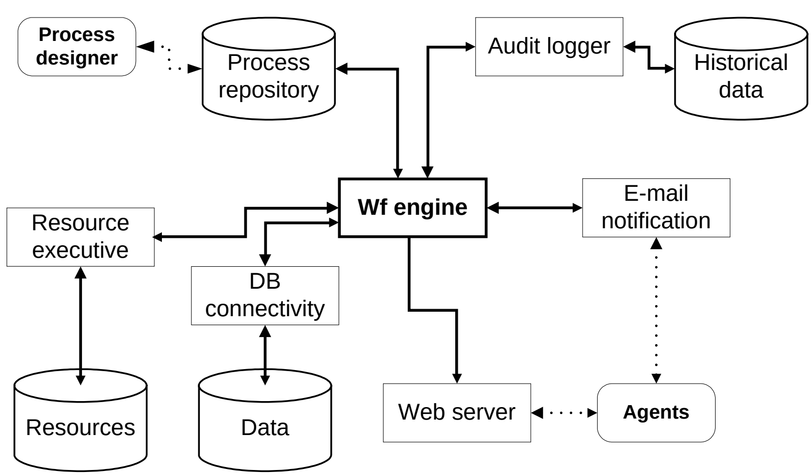
In order to support the coordinated execution of tasks by different agents, the Scheduler of the WfMS, which is a software component inside the engine of the WfMS, reads from the process repository the process model of currently active cases, and identifies the current status of execution and the next task(s) to be scheduled. The workflow engine asks the resource executive to read the organizational model and to identify the suitable agents to entrust for the execution of the task: the resource executive has to compare the skills and the authorizations of the selected agent with the minimum skills and authorizations required for the task, beyond providing workload balancing among the agents. The Scheduler also has to check for the completion of tasks within the prescribed deadlines, detecting and signaling possible anomalies and delays (Combi & Pozzi, 2003; Combi & Pozzi, 2004; Duftschmid, Miksch, & Gall, 2002).

Other components inside a WfMS are the e-mail feeder (to send and to receive e-mail messages and communications to/from agents), the Web server (to provide a html-based interface to the WfMS and to the worklists for every user), the DBMS connectivity (to access data related to current cases), and the Audit logger (to access historical data about past cases). Figure 1 depicts a general architecture for a WfMS.

WORKFLOW MANAGEMENT SYSTEMS FOR HEALTHCARE

WfMSs can be useful in coordinating any process that is "workflowable" (i.e., that can be suitably described by a process model). In order to be workflowable, a

Figure 1. A general architecture for a workflow management system (WfMS), highlighting the software components belonging to a WfMS; rectangles represent software programs; cylinders represent databases; rectangles with smoothed corners represent human executors.



process must be composed of single atomic work units, possibly assignable to different agents, and it must be repeatable and predictable. In such a way, after the process model has been identified and described, the WfMS can suitably coordinate the execution of the work units, assuring that any instance of that process will be executed according to the identified process model and within the expected deadlines.

Several applications of WfMSs can de identified in the area of healthcare, and they can be classified according to many criteria. At a very high-level view, processes managed by a WfMS can be classified according to their main goals: in the following, we assume to classify processes as self-administration processes, daily practice processes, research, and experimentation processes.

Processes for the Self-Administration

WfMSs can support typical processes which aim at self-administering the organization. For instance, a WfMS can manage the process of enrolling personnel (assigning Ids, e-mail addresses, desks, phone numbers, and so on), or of requesting authorizations to travel abroad, or of going on holiday. Generally, these processes are not strictly related to the main mission of the organization, and thus are somehow organization-independent.

Processes for the Daily Practice

WfMSs can support several processes typical of the daily practice in healthcare. Processes can be related to the treatment of patients focusing on medical aspects, and to the management of the patient dealing with organizational aspects as adhering to the recommendations from the dietician or billing the patient at discharge time.

Among some of the typical medical processes, we can consider the decision process of the physician to formulate the correct diagnosis for the patient under examination: in fact, some guidelines are defined, and can enable the physician to achieve the correct diagnosis in the shortest time, and in the less-expensive way possible. Another process can be that of adhering to the clinical guidelines specified for the identified diagnosis: in fact, after a diagnosis has been formulated, the treatment of the patient should adhere as much as possible to a standardized treatment, in order to achieve the best result in the shortest time, and to ensure that the patient has been treated according to the top-most and up-to-date criteria. These can also avoid some legal issues that could be attributed to malpractice.

As a more general process of the daily practice, we can consider the billing procedure which must be started immediately as one patient is discharged, in order to obtain the payment from the patient as soon as possible, and to allow him/her to start the reimbursement procedure from the insurance company. This process may require to collect data from different sources and from different departments, to analyze them, and to produce the final bill.

Processes for the Clinical Research

WfMSs can also support several processes for the clinical and medical research. For instance, processes can be related to the validation of treatments for patients suffering from a particular disease. In order to compare results from different experimenting centers in a uniform way, all the patients must be treated according to a common protocol, and data must be collected according to a common standard. The coordination by a WfMS is needed both to achieve the correct diagnosis and to execute the correct treatments as described by the protocol.

Moreover, WfMS could suitably support the experimental application of clinical guidelines. Clinical guidelines are recommendations, usually specified by scientific/clinical committees or associations, and based on a wide consensus from specialists, on the appropriate treatment/prevention and care of patients with specific diseases. As an example, we may consider a fragment of a guideline for stroke prevention (Inzitari, 2000):

For the treatment of endocranial hypertension, mannitol should be administered (0.25 to 0.5 g/kg for 4 hours) for patients with elevated endocranial hypertension; forthermore, glycerol should usually be administered by parenteral route (250 mL of 10% glycerol in 30-60 min every 6 hours); during therapy with glycerol, haematology should be systematically controlled since it may induce haemolysis.

According to this scenario, WfMSs could help physicians and nurses to properly manage the suitable sequence of therapeutic actions on a given patient, as well as to assign the clinical tasks to the personnel of the health organization having the right skills and roles. Moreover, WfMSs would manage possible repetitions of actions (e.g., *every 6 hours*) and exceptional conditions which do not allow the staff to follow the clinical guideline (e.g., how to deal with possible allergies of the patient to the administered drug).

FUTURE TRENDS

WfMSs still have some lacks that hurdle the widespread adoption of WfMSs themselves to manage extremely critical processes in application domains featuring such a great variability as the medical domain does.

One of the key issues to assure wide diffusion of WfMSs in managing medical processes for healthcare is that of achieving a common standard to share information (data, but mostly process definitions) among the different centers involved both in the daily practice and in research and experimentation activities. In fact, every WfMS has its own proprietary language to define processes, and no portability is assured for process definitions among different vendors: thus, the same process must be implemented in several languages, in order to enable users of different WfMSs to run the same process. Instead, the adoption of a common process definition language, like X-PDL standard language recommended by the Workflow Management Coalition, will greatly increase the ease of portability (refer to WfM Coalition Web site).

Another key issue in the adoption of WfMSs in healthcare is the need of suitably managing exceptions, which may occur during case execution (Casati, Ceri, Paraboschi, & Pozzi, 1999; Combi, Daniel, & Pozzi, 2006). In fact, the sudden worsening in the conditions of a patient cannot be neglected and—very often—may require a prompt intervention. These exceptions may lead to an immediate rescue surgery of an admitted patient or to the immediate admission to a hospital for an out-patient. Moreover, the hospital where the out-patient is admitted, due to the sudden worsening of his/her conditions can be a different hospital from the one that was taking care of that patient: the use of a treatment protocol commonly shared by the two hospitals could be very useful. The definition of such expected exceptions inside the process model will thus greatly improve and benefit the usage of a WfMS in the daily practice, as well as in the clinical research.

One more key issue, which, however, does not cover the area of WfMSs only, but spreads over information systems and decision support system, is the correct management of temporal information: in no other application domain like in healthcare a suitable management of temporal information is extremely useful. Temporal representation and temporal reasoning are some of the leading research activities in the area of temporal information management with an immediate outfall on WfMSs (Bettini, Wang, & Jajodia, 2002). Temporal representation enables one to suitably and correctly represent data over time; temporal reasoning enables one to correctly infer knowledge (including diagnosis, prognostic indices, and therapy administration) from temporal series, and from data collected in the past (Adlassnig, Combi, Das, Keravnou, & Pozzi, 2006). Both representation and reasoning over temporal information have immediate influence on modeling of decisional activities in DSSs and WfMSs.

CONCLUSION

Workflow Management Systems (WfMSs) are gaining more and more popularity and relevance, as they appear suitable to model both generic processes and specific healthcare processes. In healthcare, WfMSs can help physicians in their daily practice with well-defined treatment and diagnosing protocols. WfMSs can also help physicians in their clinical research and experimentations, especially when different research centers are involved: in fact, WfMSs provide a standardized definition of the process to be followed by the different centers involved in the validation of new treatments and therapies.

With particular relevance to healthcare, one of the major requests and expectations from WfMSs is related to the adoption of a standard language to describe the processes, while at present, no common process definition language exists.

REFERENCES

Adlassnig, K.P., Combi, C., Das, A.K., Keravnou, E., & Pozzi, G. (2006). Temporal representation and reasoning in medicine: Research directions and challenges. *Artificial Intelligence in Medicine, 38*(2), 101–113.

Bettini, C., Wang, S. X., & Jajodia, S. (2002). Temporal reasoning in workflow systems. *Distributed and Parallel Databases, 11*(3), 269–306.

Casati, F., Ceri S., Paraboschi, S., & Pozzi, G. (1999). Specification and implementation of exceptions in workflow management systems. *ACM Trans. Database Syst., 24*(3), 405-451.

Combi, C., Daniel, F., & Pozzi, G. A. (2006). Portable approach to exception handling in workflow management systems. In R. Meersman, Z. Tari (Eds.), *Cooperative Information Systems* (pp. 201–218). Berlin, LNCS: Springer Verlag.

Combi, C., & Pozzi, G. (2003). Temporal conceptual modeling of workflows. In I.-Y. Song, S. W. Liddle, T. Wang Ling, P. Scheuermann (Eds.), *Conceptual Modeling ER '03* (pp. 59–76). Berlin, LNCS: Springer Verlag.

Combi, C., & Pozzi, G. (2004). Architectures for a temporal workflow management system. In H. Haddad, A. Omicini, R. L. Wainwright, L. M. Liebrock (Eds.), *Symposium on Applied Computing* (pp. 659–666). New York: ACM Press.

Duftschmid, G., Miksch, S., & Gall, W. (2002). Verification of temporal scheduling constraints in clinical practice guidelines. *Artificial Intelligence in Medicine, 25*(2), 93–121.

Grefen, P., Pernici, B., & Sanchez, G. (1999). *Workflow management systems: The WIDE approach*. Dordrecht: Kluwer Academic Press.

Inzitari, D. (2000). The Italian guidelines for stroke prevention: The stroke prevention and educational awareness diffusion (SPREAD) collaboration. *Neurological Science, 21*(1), 5–12.

KEY TERMS

Data Model: A formal description of data managed by a business process. In most cases, these data are stored via a Database Management System (DBMS), and are also referenced by an Information System (IS) and, possibly, by a Decision Support Systems (DSS)

Exception: An event that may occur during the execution of a business process, and can—if the exception is expected and has suitably been modeled by the workflow designer—force the process to deviate from its normal flow of execution.

Organizational Model: A formal description of an organization, including agents and resources. The organizational model is used by a WfMS to assign activities to agents.

Process/Business Process: Any activity executed within an organization, requiring the coordinated execution of simple atomic activities (task) by executing agents. A business process is described by a workflow.

Process Model: A formal description of a business process. The definition is performed via a process definition language (PDL), which in most cases is WfMS-dependent.

Workflow Management Coalition (WfMC): A nonprofit organization collecting researchers, developers, users, and consultants in the area of business process automation.

Workflow Management System (WfMS): A software systems that helps to execute, automate, and enact business processes described by workflows.

X-PDL: A language, defined and standardized by the Workflow Management Coalition (WfMC) to describe business processes. X-PDL is vendor-independent, and aims at becoming a standard way for exchanging process definitions among WfMSs from different vendors.

Index

B

C

D

E

I

J

K

L

O

P

Q

R

S

T

U

V

W

X